Deep learning for detecting financial statement fraud

Patricia Craja, Alisa Kim, Stefan Lessmann

Decision Support Systems, 2020.

叶鑫 2021/09/30

Background

- Previous studies have examined various quantitative financial and linguistic factors as indicators of financial irregularities.

Study	Data (fraud / no fraud)	Features	Classifiers	Used metrics
Hajek and Henriques [26]	311/311	FIN+ LING	BBN(90.3), DTNB(89.5), RF(87. (78.0),MLP(77.9), AB(77.3), LR(	Acc, TPR, TNR, MC, F-score, AUC
Kim et al. [34]	788/2156	FIN	LR (88.4), SVM (87.7), BBN (82	Acc, TPR, G-mean, Cost Matrices
Goel and Uzuner [25]	180/180	LING+POS tags	SVM(81.8)	Acc,TPR,FPR,Precision,F-score
Purda and Skillicorn [58]	1407/4708	TXT (BOW), top 200 RF words	SVM (AUC 89.0)	AUC, Fraud Probability
Throckmorton et al. [68]	41/1531	FIN+ LING from Conference Calls	GLRT (AUC 81.0)	AUC
Goel and Gangolly [23]	405/622	LING	χ^2 statistics	χ^2 statistics
Dechow et al. [15]	293/79358	FIN	LR(63.7)	Acc, TPR, FPR, FNR, min F-Score
Humpherys et al. [29]	101/101	LING	C4.5 (67.3), NB (67.3), SVM (6	Acc, Precision, Recall, F-score
Glancy and Yadav [22]	11/20	TXT (BOW)	hierarchical clustering (83.9)	TP, TN, FP, FN, *p*-value
Perols [54]	51/15934	FIN	SVM(MC 0.0025), LR(0.0026), (	Fraud Probability and MC
Cecchini et al. [12]	61/61	LING	SVM (82.0)	AUC, TPR, FPR, FNR
Goel et al. [24]	126/622	LING+TXT (BOW)	SVM(89.5), NB(55.28%)	Acc, TPR, FPR, Precision, F-score
Lin et al. [42]	127/447	FIN	DNN (92.8), CART (90.3), LR (8	Acc, FPR, FNR, MC
Ravisankar et al. [61]	101/101	FIN	PNN (98.1), GP (94.1), GMDH (	Acc, TPR, TNR, AUC

LING: linguistic data (word category frequency counts, readability, complexity scores, etc.)
BOW: bag-of-words
POS : part of speech tags (nouns, verbs, adjectives)

Background

- Only Hajek and Henriques combined linguistic features with financial data and found that it is possible to enhance the performance through the inclusion of linguistic data.

- Furthermore, No fraud-related research has focused on the application of state-of-the-art deep learning (DL) models for textual feature extraction.

- Additionally, most previous studies neglected model interpretability, which is crucial to support auditors during client selection or audit planning.

Motivation

- The main focus of the paper is textual data processing. We introduce a novel DL method called hierarchical attention network (HAN) to extract text features from MD&A of annual reports and combine with the financial data.

- First, HAN reflects the structured hierarchy of documents, which previous approaches were unable to capture.

- Second, the HAN model embodies two different attention mechanisms at the word and sentence level, which provides “red-flag” sentences to determine whether further investigation of a specific annual report is required.

Research question

1. Does the novel combination of financial and text data (FIN+TXT) represent a more informative data type for fraud detection as compared to using FIN or TXT in isolation?

2. Can a state-of-the-art DL model(HAN) outperform the bag-of-words (BOW) approach for textual feature extraction in combination with quantitative financial features?

3. Can the proposed DL model assist in interpreting textual features signaling fraud and provide "red-flag" indicators to support the decision-making of auditors?

Research Contents

1. We select an array of classification models for detecting fraud based on different combinations of data, considering techniques including LR, SVM, RF, XGB, ANN and hierarchical attention network (HAN) .

2. All selected models are trained on five different combinations of data: financial indicators (FIN), linguistic features (LING) of an MD&A text, the full text of an MD&A (TXT).

3. We compare the predictive performance of the models with AUC, Sensitivity, F1-score, and F2-scores.

4. Following RQ 3, the HAN method provides words and sentences considered as signaling tools (red-flags) to detect financial fraud and guide the audit process.

Research Conclusion & Contribution

- 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.

- The paper bridges the gap between model’s performance and interpretability.

- The proposed method exhibits superior predictive performance and allows the identification of early warning indicators (red-flags) on both the word- and sentence-level for the facilitation of the audit process.

Research Data & Variables

- Data set: US companies' annual financial reports (10-K filings), from the EDGAR database of the SEC‟s website and quantitative financial data, from the Compustat database.

- Time: Fraud identified instances between the year 1995 and 2016.

- Amount: Using **undersampling** to balanced the data set, consisting of 1163 reports, out of which 201 are fraudulent, and 962 are non-fraudulent annual reports.

- Linguistic variables: features extracted from **the MD&A section** in 10-K filings, like sentiment, the average length of sentence, the proportion of compound words, etc.

- 47 quantitative financial variables: like total assets, profitability ratios, accounts receivable and inventories as non-cash working capital drivers.

Text-based indicators: Why MD&A?

- The MD&A is especially 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.

- The textual information is not subject to the same degree of regulation as financial information, thus providing the organization’s management more opportunities when divulging textual data.

- Breiman et al. conducted analysis on infamous examples such as WorldCom and Enron, and determined that senior managers participated in, encouraged, approved, and had knowledge of the fraudulent activities in most cases.

Text-based indicators: L&M Dictionary

- As the Loughran and Mcdonald (L&M) sentiment word lists was developed for analyzing 10-K text, it has been broadly employed in fraud-detection research.
- The sentiment categories are: negative, positive, uncertainty, litigious, strong modal, weak modal, constraining, and complexity.
  - Positive正面词词频数
  - Negative负面词词频数
  - Polarity=(Pos-Neg)/(Pos+Neg)
  - Subjectivity=(Pos+Neg)/count(*)
- Accordingly, the L&M word lists enters this study as a benchmark to DL approaches for extracting features from the MD&A section of 10-Ks.

Text-based indicators: BOW & DL

- 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.

"John likes to watch movies, Mary likes movies too"

"John also likes to watch football games"

	also	football	games	john	likes	mary	movies	to	too	watch
s1 =	[0,	0,	0,	1,	2,	1,	2,	1,	1,	1]
s2 =	[1,	1,	1,	1,	1,	0,	0,	1,	0,	1]

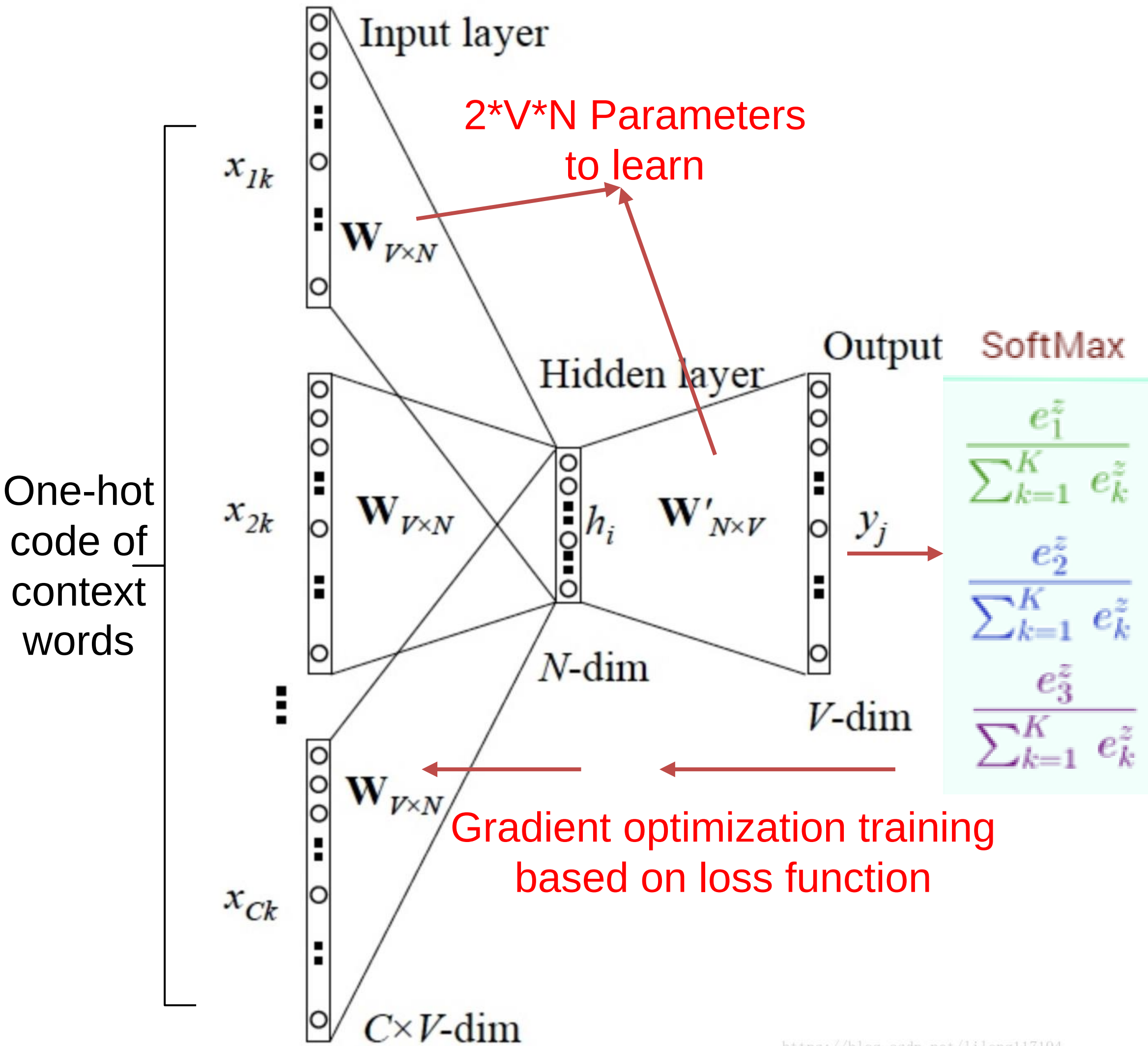
Disadvantages:

Ignore the grammar and context

High dimension and sparsity

- Textual analysis models based on DL can “learn” the **specific patterns** that underpin the text, “understand” its **meaning**, including the extraction of **contextual information** from documents.

Text-based indicators: DL-Word2vec



Word embedding
$=\text{word}_{1*v} \cdot W_{V*N}$
$= WE_{1*N}$

Advantages:

- low-dimensional dense vector
- words with similar meanings will be mapped to similar positions in the vector space

Text-based indicators: DL-Word2vec

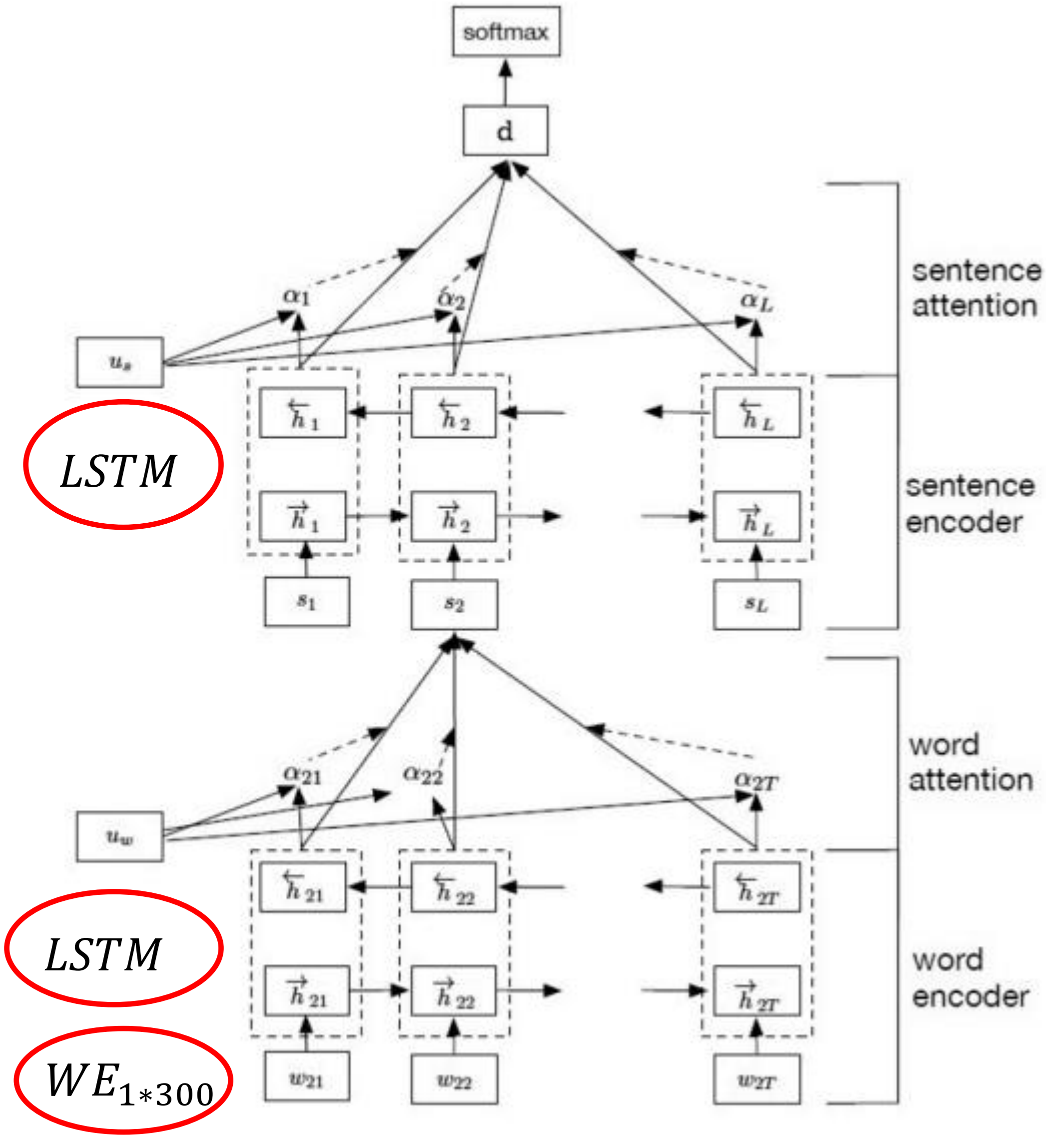
- As a result of a performance-based selection, the HAN model is built with word2vec embeddings with 300 neurons (N=300), trained on the Google News corpus, with a vocabulary size of 3 million word (V=3,000,000).

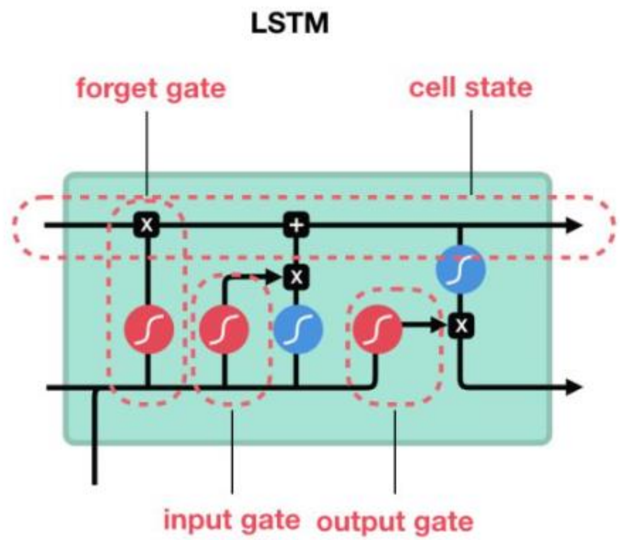
- The DL benchmark is used with the GPT-2 pre-trained embeddings from the WebText, offered by Radford et al, as they arguable constitute the current state-of-the-art language model.

- They constitute the first layer of the HAN model and allow further processing of text input within the DL architecture.

- The HAN 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).

Research Method: HAN Model



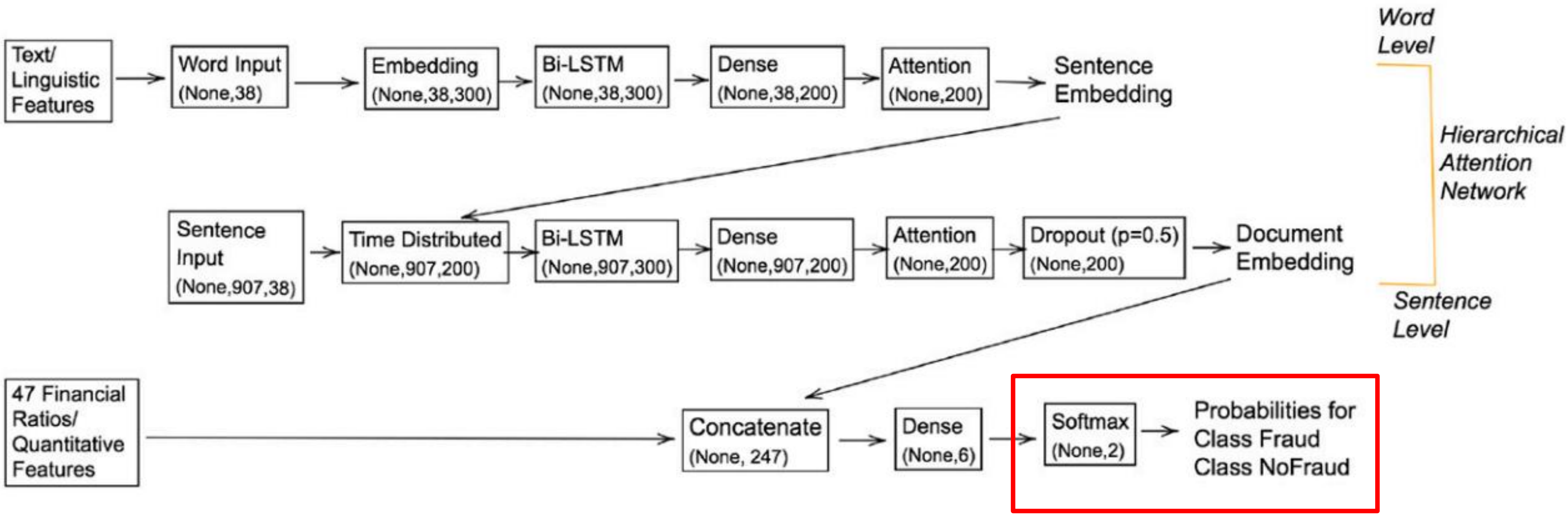


$$u_{it} = \tan h(W_w h_{it} + b_w)$$

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

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

Research Method: HAN Model



The HAN model consists of an encoder that generates relevant contexts (bidirectional LSTM) and an attention mechanism, which calculates importance weights (word attention and sentence attention).

Research Method: Evaluation metrics

- AUC
- precision = TP / (TP + FP)
- sensitivity = TP / (TP + FN)
- specificity = TN / (TN + FP)
- accuracy = (TP + TN) / (TP + TN + FP + FN)
- $F_\beta - \text{score} = (1+\beta^2) \times \dfrac{\text{precision} \times \text{sensitivity}}{(\beta^2 \times \text{precision}) + \text{sensitivity}}, \ \beta = 1, 2$

- Hajek and Henriques [26] estimated the cost of failing to detect fraudulent statements (type II error) to be twice as high as the type I error.

- So our study employs the F2-score in addition to the F1-score, as it weights sensitivity higher than precision and is, therefore, more suitable for fraud detection.

Empirical result: Answer RQ 1 and 2

- In terms of AUC and accuracy, the tree-based models RF and XGB appear to excel at predicting fraud, indicating a non-linear dependency between financial indicators and the fraud status of a report.

- XGB's high performance is noteworthy since it was not considered in prior work on fraud detection.

	AUC	Sensitivity	F1-score	F2-score	Accuracy	Delta AUC	Delta F1
Linguistics data (LING)						Comparison to FIN	
LR	0.6719	0.7000	0.3962	0.6398	0.8280	−0.0901	−0.0805
RF	**0.7713**	**0.7500**	**0.4839**	0.7302	**0.8424**	−0.0896	−0.0669
SVM	0.7406	0.7000	0.4285	0.6857	0.8280	−0.0155	−0.0340
XGB	0.7219	0.3666	0.4489	**0.8385**	0.8338	−0.1251	−0.1350
ANN	0.6782	0.6333	0.3958	0.6758	0.6676	−0.0782	−0.0605
Finance data + Linguistics data (FIN + LING)						Comparison to FIN	
LR	0.7682	0.7666	0.4623	0.6984	0.8280	0.0062	−0.0144
RF	0.8606	0.7666	0.5197	0.7610	0.8567	−0.0003	−0.0311
SVM	0.7973	0.7166	0.4858	0.7448	0.8280	0.0567	0.0573
XGB	**0.8651**	0.8166	**0.5444**	**0.7687**	**0.8653**	0.0181	−0.0395
ANN	0.7733	**0.8333**	0.4566	0.6614	0.6590	0.0169	0.0003

*classifying all cases of the test set as non-fraudulent (majority class) is **82.81%**.*

Empirical result: Answer RQ 1 and 2

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.

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		

BOW+ TF-IDF

DL-Word2vec

TXT >LING

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	0.0939	−0.0795
RF	0.8797	0.6660	0.9550	0.7079	0.9043	0.8739	0.0191	0.1882	−0.1571
SVM	0.8902	0.7833	0.8961	0.6861	0.8784	0.8280	0.0929	0.2003	−0.2576
XGB	0.8983	0.7000	**0.9653**	**0.7500**	**0.9187**	**0.9083**	0.0332	0.2056	−0.1661
ANN	0.8911	0.7460	0.9405	0.7401	0.9055	0.9054	0.1178	0.2835	−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			

FIN+TXT >FIN+LING>FIN

*classifying all cases of the test set as non-fraudulent (majority class) is **82.81%**.*

Decision support: Red Flag

- In contrast to BOW, the HAN model considers the grammar, structure, and context of words within a sentence and of sentences within a document.

- The attention mechanisms of the HAN contributes the most in attributing the fraudulent behaviour by extracting the word and sentence attention weights defined in Equation 2 and 8:

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

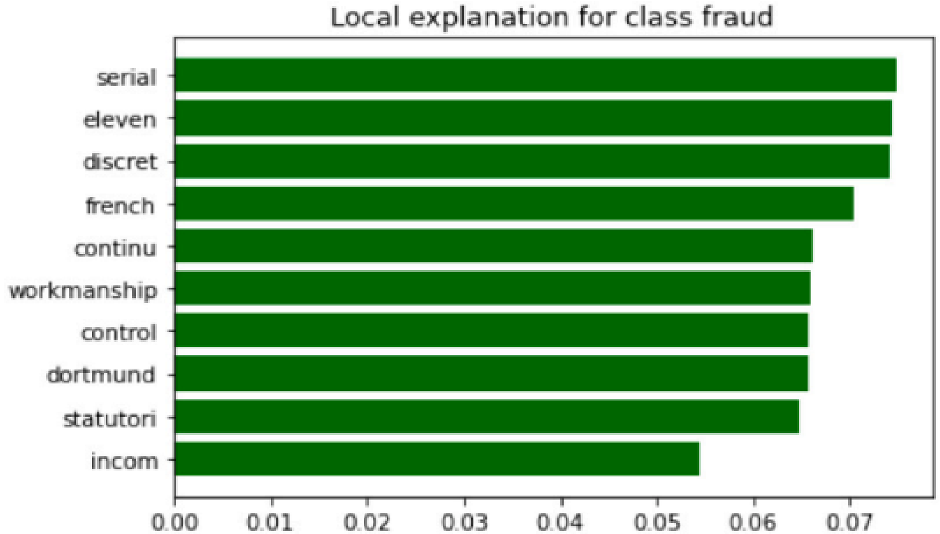


Fig. 3. Words with top weights indicating fraud from a sample MD&A.

Decision support: Red Flag

- We extract the sentence-level attention weights for 200 fraudulent reports gained as a result of prediction by HAN and filter the top 10 most important sentences per report.

- We propose to use the probability prediction of the HAN model and assign sentence weights as a two-step decision support system for auditors.

line transactions and interaction. Our customer management software applications use encryption technology to provide the security necessary to effect the secure exchange of valuable and confidential information. **Advances in computer capabilities, new discoveries in the field of cryptography or other events or developments could result in a compromise or breach of the algorithms that these applications use to protect customer transaction data. If any compromise or breach were to occur, it could seriously harm our business, financial condition and operating results. We May Not Successfully Integrate The Products, Technologies Or Businesses From, Or Realize The Intended Benefits Of Recent Acquisitions, And We May Make Future Acquisitions Or Enter Into Joint Ventures That Are Not Successful. In the future, we could acquire additional products, technologies or businesses, or enter into joint venture arrangements, for the purpose of complementing or expanding our business.** Managements negotiations of potential acquisitions or joint ventures and managements integration of acquired products, technologies or businesses, could divert managements time and resources. Future acquisitions could cause us to issue equity securities that would dilute your ownership of us, incur debt or contingent liabilities, amortize intangible assets, or write off in process research and development and other acquisition related expenses that could seriously harm our financial condition and operating results. Further, we may not be able to properly integrate acquired products, technologies or businesses, with our existing products and operations, train, retain and motivate personnel from the acquired businesses, or combine potentially different corporate cultures. **If we are unable to fully integrate acquired products, technologies or businesses, or train, retain and motivate personnel from the acquired businesses, we may not receive the intended benefits of those acquisitions, which could seriously harm our business, operating results and financial condition.** The Loss Of Any Of Our Key Personnel Or Our Failure To Attract Additional Personnel Could Seriously Harm Our Company. We rely upon the

Conclusion

- 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 utilisation of interpretable state-of-the-art technology is essential to facilitate the detection of fraud by auditors and will significantly enhance effectiveness and efficiency of audit work.

- 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.