Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 57–64

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Financial fraud detection based on the part-of-speech features of textual risk disclosures in financial reports

Hao Sun[a,b], Jianping Li[c], Xiaoqian Zhu[c]*

[a]*Institutes of Science and Development, Chinese Academy of Sciences, No.15 Beiyitiao Alley, Haidian District, Beijing 100190, China*
[b]*School of Public Policy and Management, University of Chinese Academy of Sciences, No.19A Yuquan Road, Shijingshan District, Beijing 100049, China*
[c]*School of Economics and Management, University of Chinese Academy of Sciences, No.3 Nanyitiao Alley, Zhongguancun, Haidian District, Beijing 100190, China*

Abstract

The textual risk disclosures in the annual financial reports, which discuss the companies' potential risks in the future, are rarely considered in financial fraud detection. The purpose of this study is to detect financial fraud by incorporating the textual risk disclosures. To quantify the linguistic features of textual risk disclosures, we analyze the part-of-speech (POS) for each word in the text and measure the percentage of different types of POS words. Based on the textual risk disclosures of 8999 firm-year financial reports for U.S. energy companies from 2006 to 2019, the empirical results corroborate that the POS features can provide significant detective ability for financial fraud, and can also improve the financial fraud detection performance based on commonly used financial variables. This study is helpful for investors and regulators to understand the role of textual risk disclosures in financial reports and provide theoretical guidance for the integration of textual information in financial fraud detection.

Keywords: Financial fraud detection; Textual risk disclosures; Part-of-speech; Linguistic feature; Machine learning

1. Introduction

Recently, the continuously increasing financial fraud incidents have negatively impacted the trust between companies, gatekeepers, and market participants, and also cause a significant threat to the efficiency of financial markets [1]. Defined in the Fraud risk management Guide (2016) by the Committee of Sponsoring Organizations (COSO) and the Association of Certified Fraud Examiners (ACFE): "Fraud is any intentional act or omission designed

* Corresponding author. *E-mail address:* zxq@ucas.ac.cn

1877-0509
10.1016/j.procs.2023.07.009

to deceive others, resulting in the victim suffering a loss and/or the perpetrator achieving a gain". To protect investors, creditors, and regulators from the sustained losses caused by financial fraud, it is crucial to construct an efficient financial fraud detection model.

To construct an accurate fraud detection model, researchers explore the detective ability of various types of data [2,3]. From the very beginning, the structured financial data in the financial statement are commonly used in financial fraud detection. For instance, Dechow et al [2] construct financial ratios from financial statements and corroborate that the financial ratios are associated with financial fraud, such as the percentage change in total assets, and the discrepancy between growth measured in accounting and real growth. Johnson et al. [4] find that companies, which undergo decelerating growth in earnings per share, are more likely to be financial fraud in the future. Bao et al. [5] compile 28 raw financial items and find the financial statement information can be significantly helpful in financial fraud detection.

In recent years, a large body of studies find that unstructured textual information can also be helpful to detect financial fraud [6–8]. As a complement to common financial data, the textual information can describe the situation of companies more comprehensively [9,10]. For example, Hoberg and Lewis [7] detect financial fraud based on the Management discussion and analysis (MD&A) in the financial reports, and construct the topic model to analyze the topics that fraudulent companies are more or less likely to disclose. Purda and Skillicorn [6] construct a model based on the individual word in the MD&A of financial reports, and the results demonstrate that they can achieve better performance than common financial statement data. Dong et al. [11] adopt the systemic functional linguistics (SFL) theory to analyze the textual information in the financial social media platforms, and find social information can improve the detection performance of common financial variables. Brown et al. [12] use the topic model to derive the thematic content of textual information in financial reports and find that the textual data can provide the incremental ability for financial fraud detection. However, these types of textual information mainly discuss the companies' business currently, and rarely outlook the development in the future.

According to the requirements of regulators, listed companies are usually required to analyze the potential risks based on current operating conditions and disclose them in the financial reports in the form of texts. For example, the Securities and Exchange Commission (SEC) required the U.S. public companies to newly create "Risk Factor" section in the annual financial reports Form 10-K, which discloses the potential factors that expose the company to risk [13]. Compared with social media or MD&A in financial reports, the textual risk disclosures can more directly and portray the company's future risks [13–15]. In addition, existing research has found that this section accounts for an increasingly large portion of the overall financial report [16], and provides a true and effective picture of the risks that the company will face in the future [17]. However, few studies have explored the role of the textual risk disclosures in financial fraud detection.

The purpose of this study is to investigate whether the textual risk disclosures in financial reports can be helpful for financial fraud detection. To extract information from the textual risk disclosures, we measure the linguistic features by labeling and quantifying the part-of-speech (POS) of each word in texts, such as Adjective, Verb, and Noun. Based on the quantified POS features as detective indicators, four common machine learning methods are constructed for financial fraud detection, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost [5,18]. The results demonstrate that the POS features derived from textual risk disclosures can be significantly helpful for financial fraud detection, and can improve the detection performance based on common financial variables.

This study makes several important contributions. First, we expand the types of textual data used in existing financial fraud detection research, by incorporating the textual risk disclosures in the financial reports into financial fraud detection. The textual data used in existing studies of financial fraud detection mainly include MD&A, social media, etc.[11,12,19], which focuses on explaining and analyzing the company's current financial and business conditions. Differently, the textual risk disclosures in a company's financial report, introduced in this study, provide a more direct and forward-looking description of the risks that the company may face in the future. Second, we extend the methods to analyze the linguistic features of financial textual information, by tagging and quantifying the POS of each word in texts. When analyzing the linguistic features of text data, existing studies mainly consider the sentiment, readability, length et al[9,20]. but rarely consider the POS features of texts.

2. Method

2.1. Linguistic features analysis

To analyze the linguistic features of textual risk disclosures, we measure the percentage of different types of POS words in risk disclosures for every company. However, it is time-consuming to identify the POS tag of words by human judgment. In recent years, the common method of part-of-speech tagging is to automatically identify the POS of words in a sentence based on machine learning methods [21,22]. Thus, in this paper, we adopt the Nature Language Process tool "spaCy" to automatedly label each word in textual risk disclosures with a POS tag. Table 1 shows the description of different categories of the POS tags, including Adjective, Adposition, Adverb, et al.

Table 1. Description of POS tags

POS tags	Definition	Examples
Adjective	Adjectives are words that typically modify nouns and specify their properties or attributes	enconomic, lower, higher
Adposition	Adposition is a cover term for prepositions and postpositions	in, of, during
Adverb	Adverbs are words that typically modify verbs for such categories as time, place, direction or manner.	very, well, strong
Auxiliary	An auxiliary is a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality.	has, should ,will
Coordinating Conjunction	A coordinating conjunction is a word that links words or larger constituents without syntactically subordinating one to the other and expresses a semantic relationship between them.	and, or, but
Determiner	Determiners are words that modify nouns or noun phrases and express the reference of the noun phrase in context.	a, an, the
Interjection	An interjection is a word that is used most often as an exclamation or part of an exclamation.	bravo, ouch, hello
Noun	Nouns are a part of speech typically denoting a person, place, thing, animal or idea.	filing, risk, number
Numeral	A numeral is a word, functioning most typically as a determiner, adjective or pronoun, that expresses a number and a relation to the number, such as quantity, sequence, frequency or fraction.	One, two ,three
Particle	Particles are function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech	Not, 's
Pronoun	Pronouns are words that substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context.	Our, they, myself
Proper Noun	A proper noun is a noun (or nominal content word) that is the name (or part of the name) of a specific individual, place, or object.	Apple, Federal, London
Punctuation	Punctuation marks are non-alphabetical characters and character groups used in many languages to delimit linguistic units in printed text.	? . "
Subordinating Conjunction	A subordinating conjunction is a conjunction that links constructions by making one of them a constituent of the other.	if, while
Symbol	A symbol is a word-like entity that differs from ordinary words by form, function, or both.	+, -, =
Verb	A verb is a member of the syntactic class of words that typically signal events and actions, can constitute a minimal predicate in a clause, and govern the number and types of other constituents which may occur in the clause	emergy, increase, encounter

After labeling the words in textual risk disclosures according to the POS tags, the number of words for each type of POS tag is counted. Then, the textual risk disclosures for each company can be quantified as a vector, where each dimension corresponds to a POS tag and the value is the percentage of that type of word.

2.2. Financial fraud detection models and evaluation metrics

Based on the linguistic features extracted from risk disclosure as predictors, this study constructs four common machine learning models for financial fraud detection [5,18,23,24,25], including Logistic regression (LR), Support vector machines (SVM), Artificial neural network (ANN), Random forests (RF), and XGBoost. LR assumes a logit relation between the predictors and dichotomy financial distress and the L1 regularization on the coefficients can prevent overfitting [5]. SVM is a generalized linear model to find an optimal hyperplane, which maximizes the interval between the support vectors. SVM can also handle nonlinear relationships based on nonlinear kernel functions, which transform the samples to a higher dimensional space [18, 26]. Random forests and XGBoost are ensemble models based on multiple decision tree methods, and adopt Bagging and Boosting strategy methods, respectively [23].

To evaluate the performance of financial fraud detection models, four commonly used metrics are adopted, including Overall accuracy, Type I accuracy, Type II accuracy, and AUC (Area under the Receiver Operating Characteristic)[5,18], which respectively indicate the percentage of all samples, financial fraud samples, and non-fraud samples are correctly classified by the prediction model, as defined in Equations (1)-(3).

$$Overall\ accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$Type\ I\ accuracy = \frac{TP}{TP + FN} \tag{2}$$

$$Type\ II\ accuracy = \frac{TN}{FP + TN} \tag{3}$$

Where TP (True Positive) denotes the number of financial fraud samples correctly classified as fraud, FN (False Negative) denotes the number of financial fraud samples misclassified as non-fraud, TN (True Negative) is the number of non-fraud samples correctly as non-fraud, and FP (False Positive) is the number of non-fraud samples misclassified as fraud. The AUC metric is a measure of overall prediction performance, which is calculated by the area under the ROC (Receiver Operating Characteristic) curve. The value of AUC ranges from 0 to 1, and the closer AUC is to 1, the better prediction ability of the model.

3. Empirical results

The empirical study uses the textual risk disclosures, in financial reports of U.S. listed companies, to investigate the financial fraud ability of risk disclosures and compare the prediction ability with common financial variables.

3.1. Data description

The empirical study is based on the U.S. publicly listed companies in the energy industry, in which industry the number of financial fraud incidents has continued to increase in recent years. Consistent with Wei et al. [15], the energy companies are selected with the standard industrial classification (SIC) code. Since the SEC mandated the companies to include the “Risk Factor” section in the financial reports in 2005, U.S. companies began to disclose the risk disclosures in 2006. Thus, we collect the empirical data between the period range 2006 to 2019. Based on a crawler program, the Form 10-K filings released by energy companies are collected from the Electronic Data Gathering and Retrieval (EDGAR) database on SEC website. After extracting the textual risk disclosures in the “Risk Factor” section from Form 10-K fillings, we collect 11085 Form 10-K filings from 1321 U.S. energy companies from 2006 to 2019.

To explore whether the textual information can be helpful for fraud detection, the commonly used financial variables are constructed as the benchmark. Comparing the detection performance with the financial ratios proposed by previous study [2], Bao et al. [5], find that the 28 raw financial variables constructed by themselves can detect financial fraud more accurate. Thus, we collect the 28 financial variables, presented in Table 2, as the baseline

variables of fraud detection. After removing the observations with missing financial information, we obtain the remaining 8999 firm-year observations.

Table 2. Description of financial variables

No.	Variable	No.	Variable
1	Common shares outstanding	15	Assets, total
2	Current assets, total	16	Long-term debt issuance
3	Sale of common and preferred stock	17	Income before extraordinary items
4	Property, plant and equipment, total	18	Long-term debt, total
5	Account payable, trade	19	Interest and related expense, total
6	Cash and short-term investments	20	Income taxes, total
7	Price close, annual, fiscal	21	Current liabilities, total
8	Retained earnings	22	Sales/turnover (net)
9	Inventories, total	23	Income taxes payable
10	Common/ordinary equity, total	24	Investment and advances, other
11	Debt in current liabilities, total	25	Liabilities, total
12	Depreciation and amortization	26	Short-term investments, total
13	Receivables, total	27	Net income (loss)
14	Cost of goods sold	28	Preferred/preference stock (capital), total

Following prior research [5], we identify the financial fraud companies according to the fraud samples from two commonly used fraud databases: the SEC's Accounting and Auditing Enforcement Releases (AAERs) database provided by The University of California-Berkeley Center, and the Sandford Securities Class Action Clearinghouse (SCAC) database of securities class action lawsuits filed. Finally, 76 fraud firm-year observations are collected, and the remaining 8923 firm-years observations as the sample of non-fraud observations. Table 3 summarizes the fraud and non-fraud companies' distribution by year.

Table 3. Fraud and non-fraud companies' distribution by year

Year	Fraud companies	Non-fraud companies	Year	Fraud companies	Non-fraud companies
2006	2	608	2013	7	683
2007	2	626	2014	3	666
2008	3	664	2015	6	634
2009	2	683	2016	7	595
2010	6	680	2017	9	593
2011	5	672	2018	8	587
2012	5	684	2019	11	548

3.2. Linguistic features extraction

Based on the linguistic features introduced in section 2.1, we calculate the percentage of each type of POS tag in textual risk disclosures. Table 4 shows the statistical results of different types of POS tags. Comparing the mean values of different types of POS, we can find that the mean percentage of Noun words is 26.43%, which is higher than other types of POS words. This result suggests that the Noun words take up the main content in the textual risk disclosures. Besides, Verb, Adposition, Determiner, Punctuation, and Adjective words are the other 4 types of POS words that account for more than 9% of the content in the text on average. The percentages of the remaining types of POS words are below 4% on average, such as Adverb, Auxiliary, and Particle.

Table 4. Statistical results of POS features

POS features	Max (%)	Min (%)	Mean (%)	Median (%)	Std (%)
Adjective	15.09	2.90	9.24	9.28	0.93
Adposition	17.39	4.34	10.80	10.78	0.74
Adverb	5.00	0.00	2.39	2.38	0.36
Auxiliary	5.58	0.00	2.92	2.88	0.43
Coordinating Conjunction	0.00	0.00	0.00	0.00	0.00
Determiner	16.61	0.75	10.44	10.50	1.22
Interjection	0.25	0.00	0.01	0.00	0.02
Noun	33.36	10.53	26.43	26.44	1.77
Numeral	8.80	0.00	0.87	0.78	0.55
Particle	8.26	0.00	2.73	2.63	0.58
Pronoun	4.71	0.00	1.76	1.95	0.83
Proper Noun	44.36	0.00	3.87	3.15	2.47
Punctuation	20.29	6.42	9.35	9.30	0.84
Subordinating Conjunction	4.57	0.00	1.53	1.52	0.31
Symbol	1.88	0.00	0.11	0.08	0.12
Verb	17.77	1.50	11.83	11.87	0.99

3.3. Fraud detection performance using the textual risk disclosures

This section investigates the detective ability of textual risk disclosures in companies' annual financial reports. Based on the POS features extracted from the textual risk disclosures, we compare the financial fraud detection performance of POS features and common financial variables, and then analyze whether the POS features can provide incremental information relative to financial variables.

In the process of implementing financial fraud detection, four common models shown in section 2.3 are constructed, including LR, SVM, RF, and XGBoost [5,18,23,24,25]. Specifically, the regularization parameter L1 in LR and SVM is 5, the kernel function in SVM is Gaussian kernel function, and the number of base decision trees in RF and XGBoost is 100. The dataset is randomly divided into 80% as training data and 20% as testing data. Since the hyperparameters in each model need tuning, 20% of training data are divided as the validation data, and the Grid Search method is used to search for the optimal hyperparameters for each model. The prediction performance could be impacted negatively by the imbalanced ratio of financial fraud samples versus non-fraud samples. To mitigate the negative impact of class imbalance problem, we adopt the Cost Sensitive learning method to assign higher penalty weights to the classification cost for financial fraud samples relative to normal samples, and the optimal penalty weights are determined by the Grid Search method. Finally, the prediction performance in the testing data of five models with different predictive information can be evaluated by the average of 5-fold cross-validation.

Table 5 summarizes the detection performance of financial variables, POS features, and the combination of them, respectively. According to the results of common financial variables, we can find that the AUC of XGBoost is 70.81%, which is higher than other models, and its Overall accuracy, Type I accuracy and Type II accuracy are 68.01%, 68.75%, 68.00%, respectively. In the results of fraud detection by Bao et al. [5], who construct the 28 baseline financial variables, the detection performance of LR and SVM models are 69.00% and 62.60%, respectively. The better detection performance of the LR and SVM model in our study demonstrates that the selection of the financial variables as the benchmarks is reliable. The difference in the AUC value might be caused by different datasets, where the datasets of Bao et al.[5] is all industries and our dataset is based on the energy industry.

Table 5. Financial fraud detection performance based on POS features of textual risk disclosures

Prediction variables	Prediction models	Type I accuracy (%)	Type II accuracy (%)	Overall accuracy (%)	AUC (%)
Financial variables	LR	65.00	64.68	64.68	69.50
	SVM	66.25	67.53	67.52	69.86
	RF	62.50	62.64	62.64	67.35
	XGBoost	**68.75**	**68.00**	**68.01**	**70.81**
POS features of textual risk disclosures	LR	65.00	64.76	64.77	68.34
	SVM	67.50	**67.52**	**67.52**	69.94
	RF	63.75	64.81	64.80	69.07
	XGBoost	**67.50**	66.46	66.47	**71.48**
Financial variables + POS features of textual risk disclosures	LR	67.50	66.81	66.82	72.51
	SVM	65.00	64.07	64.08	70.40
	RF	67.50	67.37	67.37	72.47
	XGBoost	**67.50**	**68.45**	**68.45**	**73.89**

Note:Numbers in bold-face indicate the best performance for each metric.

According to the results of POS features, we can find the best AUC XGBoost is 71.48%, which is higher than other models, and its Overall accuracy, Type I accuracy and Type II accuracy are 66.47%, 67.50%, 66.46%, respectively. Compared with the performance of financial variables, the POS features can better detect financial fraud. Moreover, when adding the POS features based on financial variables, the best AUC is 73.89% by XGBoost, and its Overall accuracy, Type I accuracy and Type II accuracy are 68.45%, 67.50%, 68.45%, respectively. Relative to the performance of using financial variables only, the AUC improves by 3.09%. These results suggest that the POS features of textual risk disclosures can provide significant detective ability for financial fraud and can also improve the detection performance based on common financial variables.

4. Conclusion

This study uses the textual risk disclosures in financial reposts to detect financial fraud. We label and quantify the POS tag of each word in textual risk disclosures, and then construct machine learning methods with these POS features to detect financial fraud. Based on the 8999 firm-year observation of U.S. energy listed companies, the empirical results demonstrate that the POS features of textual risk disclosures can provide significant detective ability for financial fraud, and improve the performance based on common financial variables.

The findings provide a new perspective for market participants and regulators when analyzing the possibility of financial fraud. In addition to the company's financial, market, and other structured numerical data, we should also pay attention to the unstructured textual information disclosed in the financial reports. Future research can consider various types of textual information, such as letters of inquiry from regulators, and announcements of related party transactions disclosed by companies, in order to detect the financial fraud from more perspectives.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (T2293774, 71971207, 92046023), Fundamental Research Funds for the Central Universities (UCAS-E2EG0805X2, UCAS-E2ET0808X2), and MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation at the University of Chinese Academy of Sciences.

References

[1] Dan Amiram, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff and Richard Sloan. (2018) "Financial reporting fraud and other forms of misconduct: A multidisciplinary review of the literature." *Review of Accounting Studies* **23 (2)**: 732–783.

[2] Patricia M. Dechow, Weili Ge, Chad R. Larson and Richard G. Sloan. (2011) "Predicting material accounting misstatements." *Contemporary Accounting Research* **28 (1)**: 17–82.

[3] Wenzhou Hong, Xuxia Wang and Haiqi Ping. (2014) "Research on the financial report fraud detection of listed companies based on logistic regression model." *Chinese Journal of Management Science* **22 (S1)**: 351–356. (in Chinese)

[4] Shane A Johnson, Harley E Ryan and Yisong S Tian. (2008) "Managerial incentives and corporate fraud: The sources of incentives matter*." *Review of Finance* **13 (1)**: 115–145.

[5] Yang Bao, Bin Ke, Bin Li, Y. Julia Yu and Jie Zhang. (2020) "Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach." *Journal of Accounting Research* **58 (1)**: 199–235.

[6] Lynnette Purda and David Skillicorn. (2015) "Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection." *Contemporary Accounting Research* **32 (3)**: 1193–1223.

[7] Gerard Hoberg and Craig Lewis. (2017) "Do fraudulent firms produce abnormal disclosure?" *Journal of Corporate Finance* **43**: 58–85.

[8] Chandra S. Throckmorton, William J. Mayew, Mohan Venkatachalam and Leslie M. Collins. (2015) "Financial fraud detection using vocal, linguistic and financial cues." *Decision Support Systems* **74**: 78–87.

[9] Tim Loughran and Bill McDonald. (2016) "Textual analysis in accounting and finance: A survey." *Journal of Accounting Research* **54 (4)**: 1187–1230.

[10] Jianping Li, Guowen Li, Mingxi Liu, Xiaoqian Zhu and Lu Wei. (2022) "A novel text-based framework for forecasting agricultural futures using massive online news headlines." *International Journal of Forecasting* **38 (1)**: 35–50.

[11] Dong Wei, Shaoyi Liao and Zhongju Zhang. (2018) "Leveraging financial social media data for corporate fraud detection." *Journal of Management Information Systems* **35 (2)**: 461–487.

[12] Nerissa C. Brown, Rrichard M. Crowley and W. Brooke Elliott. (2020) "What are you saying? Using topic to detect financial misreporting." *Journal of Accounting Research* **58 (1)**: 237–291.

[13] Xiaoqian Zhu, Yinhui Wang and Jianping Li. (2022) "What are the driving factors behind reputation risk events: Evidence from financial institutions' textual risk disclosures." *Humanities and Social Sciences Communications* **9**: 318.

[14] Lu Wei, Guowen Li, Jianping Li and Xiaoqian Zhu. (2019) "Bank risk aggregation with forward-looking textual risk disclosures." *The North American Journal of Economics and Finance* **50**: 101016.

[15] Lu Wei, Guowen Li, Xiaoqian Zhu, Xiaolei Sun and Jianping Li. (2019) "Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures." *Energy Economics* **80**: 452–460.

[16] Travis Dyer, Mark Lang and Lorien Stice-lawrence. (2017) "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics* **64 (2–3)**: 221–245.

[17] John L. Campbell, Hsinchun Chen, Dan S. Dhaliwal, Hsinmin Lu and Logan B. Steele. (2014) "The information content of mandatory risk factor disclosures in corporate filings." *Review of Accounting Studies* **19 (1)**: 396–455.

[18] Bertomeu Jeremy, Edwige Cheynel, Eric Floyd and Wenqiang Pan. (2021) "Using machine learning to detect misstatements." *Review of Accounting Studies* **26 (2)**: 468–519.

[19] Yiyun Chen. (2019) "Forecasting financial distress of listed companies with textual content of the information disclosure:A study based MD&A in chinese annual reports." *Chinese Journal of Management Science* **27 (7)**: 23–34. (in Chinese)

[20] Dewen Liu, Weihe Gao and Liangyu Min. (2022) "The impact of readability and attractivenes on product sales—Text analysis based on movie introduction." *Chinese Journal of Management Science* **30 (6)**: 167–177. (in Chinses)

[21] Khan Wahab, Ali Daud, Khairullah Khan, Jamal Abdul Nasir, Mohammed Basheri, Naif Aljohani and Fahd Saleh Alotaibi. (2019) "Part of speech tagging in urdu: Comparison of machine and deep learning approaches." *IEEE Access* **7**: 38918–38936.

[22] Wenhao Zhu, Tengjun Yao, Wu Zhang and Baogang Wei. (2019) "Part-of-speech-based long short-term memory network for learning sentence representations." *IEEE Access* **7**: 51810–51816.

[23] Xiaoqian Zhu, Xiang Ao, Zidi Qin, Yanpeng Chang, Yang Liu, Qing He and Jianping Li. (2021) "Intelligent financial fraud detection practices in post-pandemic era: A survey." *The Innovation* **2 (4)**: 100176.

[24] Jianping Li, Yanpeng Chang, Yinhui Wang and Xiaoqian Zhu. (2023) "Tracking down financial statement fraud by analyzing the supplier-customer relationship network." *Computers & Industrial Engineering* **178**: 109118.

[25] Chenggang Li, Hongye Jia, Guanghui Zhao and Hong Fu. (2023) "Credit risk warning of listed companies based on information disclosure text: Empirical evidence from management discusion and analysis of the chinese annual report." *Chinese Journal of Management Science* **31 (2)**: 18–29. (in Chinese)

[26] Hongyan Yang and Yingjie Tian. (2022) "Application research of machine learning in food safety risk early warning and sampling inspection program." *Management Review* **11**: 315-323. (in Chinese)