Speech Emotion Recognition from Earnings Conference Calls in Predicting Corporate Financial Distress

Petr Hajek

Speech Emotion Recognition from Earnings Conference Calls in Predicting Corporate Financial Distress *

Petr Hajek[1][0000−0001−5579−1215]

Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, Pardubice, Czech Republic
petr.hajek@upce.cz

Abstract. Sentiment and emotion analysis is attracting considerable interest from researchers in the field of finance due to its capacity to provide additional insight into opinions and intentions of investors and managers. A remarkable improvement in predicting corporate financial performance has been achieved by considering textual sentiments. However, little is known about whether managerial affective states influence changes in overall corporate financial performance. To overcome this problem, we propose a deep learning architecture that uses vocal cues extracted from earnings conference calls to detect managerial emotional states and exploits these states to identify firms that could be financially distressed. Our findings provide evidence on the role of managerial emotional states in the early detection of corporate financial distress. We also show that the proposed deep learning-based prediction model outperforms state-of-the-art financial distress prediction models based solely on financial indicators.

Keywords: Speech emotion recognition · Financial distress · Deep learning · Earnings conference calls.

1 Introduction

Financial distress prediction models are widely regarded as some of the most important models in finance due to their capacity to provide early warnings to stakeholders about a firm's impending business failure. Stakeholders have suffered significant losses during recent financial crises. Their plight increases the need to reduce information asymmetries between corporate managers and other stakeholders and provide predictive models of financial distress.

Most models to date tended to focus on corporate financial indicators when predicting financial distress [1]. However, in recent years, there has been an increasing amount of literature on the role of sentiment and emotion analysis in a firm's textual documents. Indeed, the additional insights gained about investor

* Supported by the scientific research project of the Czech Sciences Foundation Grant No: 19-15498S.

and managerial opinions proved to be significant predictors of corporate financial performance. Previous research in this regard has focused on identifying sentiment in managerial communications in annual reports [2–5]. However, the linguistic tone of conference calls also turned out to be a significant predictor of abnormal financial returns [6]. It has been shown that the significance of emotional information can outweigh that of factual financial information disclosed by managers and thus indicate financial risks to a company [7]. Recent evidence also shows that nonverbal managerial communication is incrementally useful when combined with quantitative indicators in predicting corporate financial performance [8, 9]. This is explained by the dissonance (cognitive conflict) between a manager's emotional state and a firm's actual financial performance, dissonance that can indicate potential financial distress.

With the above thoughts as a basis, we assert that certain emotions detected in earnings conference calls may indicate corporate financial distress. To the best of our knowledge, this is the first work incorporating speech emotion recognition (SER) in financial distress prediction models. To address this issue, we here propose a hybrid deep learning model that exploits state-of-the-art convolutional neural network (CNN)-based SER to leverage financial distress prediction using a long short-term memory (LSTM) neural network. By combining nonverbal vocal attributes obtained from audio recordings of earnings conference calls and financial indicators from financial statements, we hypothesize that the proposed prediction model will outperform traditional models of financial distress, which are based purely on financial data. In addition, we aim to greatly increase the understanding of the role specific emotions might play in predicting corporate financial distress.

The remainder of this paper is organized as follows. Section 2 presents related literature on the use of managerial vocal cues in finance. Section 3 outlines the conceptual framework proposed for early detection of corporate financial distress. Section 4 presents the vocal and financial variables and describes our data. Section 5 describes the setting of the proposed deep learning architecture. Section 6 shows the experimental results of the proposed SER-based financial distress prediction model, and compares the model with existing approaches based on financial variables. Section 7 provides conclusions and outlines future research directions.

2 Related Literature on Using Vocal Cues in Finance

Previous studies have shown that the accuracy of financial prediction models can be significantly increased by exploiting vocal cues from earnings conference calls. A list of those studies is given in Table 1 and shows the data and vocal features used, the methods used, and the prediction problem addressed in terms of predicted variable.

Regarding the vocal cues used, previous studies mostly analyzed audio recordings from earnings conference calls using two tools, layered voice analysis (LVA) [8, 10] and Praat voice analysis [11, 12]. LVA allows the user to extract the level of

affective states from audio recordings, including different categories of cognitive states and emotional reactions. More precisely, four essential features can be obtained, namely the cognitive level, emotional level, thinking level, and global stress level. The cognitive level captures the cognitive dissonance, and the level of excitement is captured by the emotional level. Mental efforts and physical arousal are approximated using the thinking level and stress level features, respectively. Notably, abnormally high emotional levels indicate a positive affect. Mayew [8] and Price [9] reported that positive and negative affects are significant determinants of cumulative abnormal stock returns. Vocal dissonance markers proved to be useful for identifying financial misreporting [10].

Table 1. Summary of data and methods used in previous studies

Study	Method	Features	Prediction task
[8]	LVA+MLR	Positive affect, negative affect	Cumulative abnormal return
[10]	LVA+MLR	Cognitive dissonance	Financial misreporting
[13]	Praat+GLRT	Fundamental frequency, small-scale perturbations, variations of amplitude maxima, mean harmonics-to-noise ratio, proportion of voiced speech	Financial fraud detection
[14]	SPLCE+ MANOVA	Pitch and voice quality, vocal intensity, response latency, pitch slope	Identifying potentially fraudulent utterances
[9]	LVA+MLR	Emotional activity level, cognitive activity level	Cumulative abnormal return
[11]	Praat+HTML	27 vocal features including pitch, intensity, jitter, and the harmonics-to-noise ratio	Stock price volatility forecasting
[12]	Praat+LSTM	26 vocal features including pitch, intensity, jitter, and the harmonics-to-noise ratio	Stock price volatility forecasting
[15]	Praat+SVM	26 vocal features including pitch, intensity, voice, and harmonicity	Stock price volatility and price forecasting
[16]	Praat+DNN	26 vocal features including pitch, intensity, voice, and harmonicity	Stock price volatility forecasting
[18]	pyAudioAnalysis	Emotion valence, emotion arousal	Artificial intelligence readership
[17]	Praat+MDRM	26 vocal features including pitch, intensity, voice, and harmonicity	Stock price volatility forecasting
This study	CNN-based SER+LSTM	180 spectral features used to detect 8 emotional states	Corporate financial distress prediction

Legend: DNN – deep neural network, GLRT – generalized likelihood ratio test, HMTL – hierarchical transformer-based multi-task learning, LVA – layered voice analysis, MANOVA – multivariate analysis of variance, MDRM – multi-modal deep regression model, MLR – multivariate linear regression, SPLCE – structured programming for linguistic cue extraction, SVM – support vector machine.

From a different perspective, the Praat software allows researchers to quantify a wide range of acoustic features, such as pitch, intensity, jitter, shimmer, and excitation patterns. Throckmorton [13] showed that combining features across linguistic and vocal categories provided better fraud detection than those using financial indicators but only if feature selection was performed. Similarly, statements covering up fraud were reportedly higher pitched and lower in voice quality than legitimate statements [14]. Vocal features extracted using Praat also improved the prediction of stock volatility [11]. A semi-supervised multi-modal learning model was shown to be effective for stock volatility prediction [12], and a neural attentive alignment model effectively captured interdependencies across vocal and verbal modalities in another stock price volatility forecasting model [15]. In a similar manner, cross-model and inter-modal attention for deep

verbal-vocal coherence was used for modelling stock price interdependence [16]. Most recent works found that gender bias exists in multi-modal volatility prediction [17] and that company representatives adjust the way they talk when they know machines are listening [18].

To summarize the above findings, until now vocal cues have only been applied to financial fraud detection, cumulative abnormal return forecasting, and stock price volatility forecasting. Recent developments in deep learning-based SER allow us to identify the managerial emotional state in earnings conference calls with high accuracy.

3 Conceptual Framework for Predicting Financial Distress

The conceptual framework proposed in this study is depicted in Fig. 1. As reported above, earlier related studies have only considered managers' vocal features rather than the direct detection of managers' emotions expressed in earnings conference calls. To overcome this limitation, inspired by [19], we first employed a CNN-based SER model, which provided the state-of-the-art performance for the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) benchmark dataset [20]. The dataset comprises 1,440 recordings classified into eight emotional states: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. This allowed us to recognize eight different emotions. Then, audio recordings of earnings conference calls were fed to the trained SER model to obtain emotional state labels. In the next step, the extracted eight emotional features were combined with 20 financial indicators calculated from financial statement data. Finally, data for the previous four quarters were used in the LSTM model for the 1-year-ahead prediction of financial distress. The LSTM recurrent neural network was used to effectively capture high-level temporal features from sequential quarterly data to accurately predict financial distress. We discuss the proposed deep learning-based architecture in detail in Section 5.

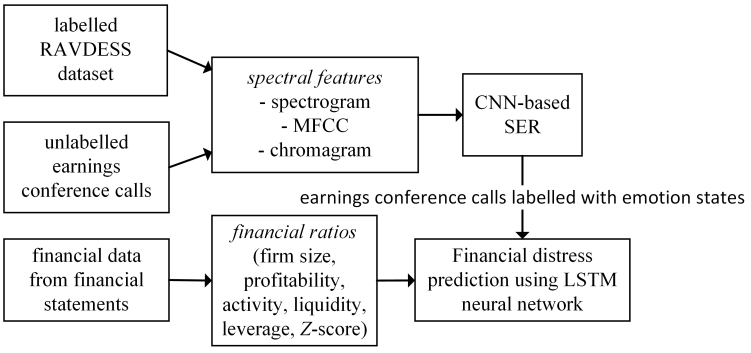

Fig. 1. Conceptual framework for financial distress prediction.

4 Data and Features

The audio file dataset consisted of 1,278 earnings conference calls collected on a quarterly basis from Q1 2010 to Q3 2021 from the EarningsCast database at https://earningscast.com/. The audio data are freely available to the public. For our sample, we considered 40 companies in the United States listed on the New York Stock Exchange (NYSE) with the largest market capitalizations. The reason we chose the earnings conference calls is because they feature business executives discussing their companies' financial information in public, thus conveying financial and voice signals simultaneously [13]. The downloaded audio files were converted to .wav files. It is worth noting that the audio recordings are provided without any segmentation or labels for speakers.

To extract features for SER, we used the Librosa audio library [21]. Specifically, we used 180 spectral features grouped into three different categories, namely, mel-frequency cepstral coefficients (MFCCs) (40 features), mel-scaled spectrograms (128 features), and chromagrams (12 features). These features simulate the way humans receive sound frequencies, with MFCCs forming a mel-frequency cepstrum and the mel scale representing the nonlinear mapping (Fourier transform) of the frequency scale. To represent pitch classes and harmony, chromagrams were obtained using short-time Fourier transform [21]. Overall, our intention was to get a rich representation of the audio recordings, which allowed us to achieve high accuracy in the CNN-based SER model. To illustrate the obtained features, Fig. 2 shows the spectral features for Adobe in Q4 2019.

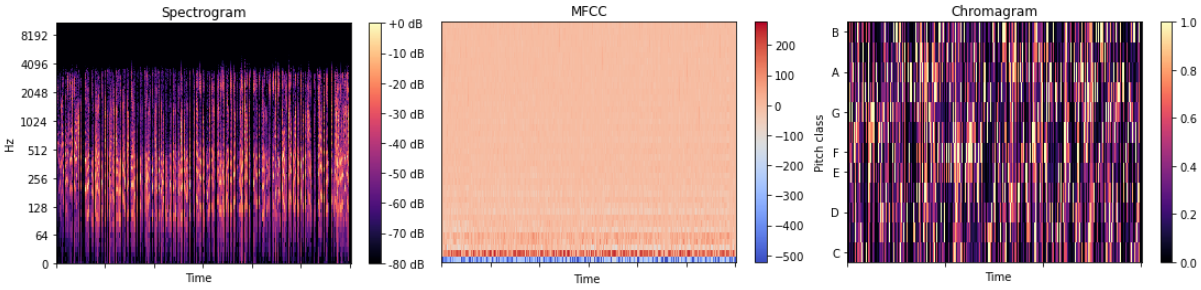

Fig. 2. Illustration of spectral features for Adobe in Q4 2019.

To obtain the matching financial features, we utilized financial statement data collected from the freely available Macrotrends database (www.macrotrends.net). In agreement with related studies on financial distress prediction [22, 23], the following financial features were included: (1) firm size (total assets, sales, cash flow, equity), (2) profitability ratios (retained earnings to total assets, return on total assets, return on equity, gross margin, operating margin), (3) activity ratios (asset turnover, inventory turnover, receivable turnover), (4) liquidity ratios (current ratio, cash ratio, working capital to total assets, operating cash flow per

share, free cash flow per share), (5) leverage ratios (equity to book value of total liabilities, total debt), and (6) overall financial performance indicator (Altman's Z-score).

Corporate financial distress is defined as a firm's inability to meet its payment obligations on debt. The Altman's model (Altman's Z-score) [24] is the most widely used model for early detection of corporate financial distress. We chose this model to categorize the companies into three classes, namely, safe, grey, and distress zones, because the model was specifically designed for companies listed on U.S. public capital markets. In addition, the model has been shown to be valid for developed markets [24]. The model is able to predict the bankruptcy of a company with high accuracy up to 2 years in advance. The Z-score model for companies with shares publicly tradeable on stock markets is given as follows:

$$Z{-}score = 1.2x_1 + 1.4x_2 + 3.3x_3 + 0.6x_4 + 1.0x_5, \tag{1}$$

where x_1 denotes working capital to total assets, x_2 is retained earnings to total assets, x_3 is return on total assets, x_4 is equity to book value of total liabilities, and x_5 is asset turnover. A Z-score of 3 or higher indicates the safe zone (a company with high probability to survive), of 1.80 to 2.99 denotes the grey zone (a company with certain financial difficulties), and of less than 1.80 indicates the distress zone (a high risk for financial distress).

Based on the above groupings for the Z-score, the companies were put into three classes, with 59 samples in the safe zone (4.6%), 329 in the grey zone (25.7%), and 890 in the distress zone (69.7%), indicating a class imbalance problem. Notably, according to [24], the high proportion of companies in the distress zone may indicate a coming financial crisis. The classes were always assigned with a 1-year lag in order to predict financial distress classes 1 year in advance. For the next set of experiments, we additionally categorized the samples into two classes based on whether there was an increase or decrease in the Z-score during the following year. According to the trend of the Z-score, 615 samples were put into the upward class and 663 samples into the downward class. Data for 2010 to 2016 were used as training data, and data for 2017 to 2020 were used to test the performance of the prediction model.

5 Deep Learning Model for Predicting Financial Distress

The deep learning architecture we proposed is depicted in Fig. 3. In it, we used the CNN model for the classification of 8 emotions based on 180 spectral features. As shown in Fig. 3, the model includes one-dimensional convolutional layers. The proposed CNN model is inspired by the architecture developed by [19] and modified for eight emotion classes. Based on the work of [19], the first and second convolutional layers comprised 256 and 128 filters, respectively, with a kernel size of $k = 5$ and a stride step of 1. Next, the max-pooling layer was included with a window size set to 8, followed by another convolutional layer with 128 filters ($k = 5$ and stride = 1). Flattening allowed us to continue with the connection of a fully connected layer, followed by a dropout layer (with a dropout rate of 0.2)

and a softmax layer with the number of neurons corresponding to the number of predicted emotional classes. Training was performed using the Adam optimizer with a learning rate of 0.0001 and cross-entropy loss as the fitness function. Five-fold cross-validation was used for model evaluation. After the SER model was trained on the RAVDESS dataset, it was possible to use it to label the audio recordings of earnings conference calls based on the 180 spectral features.

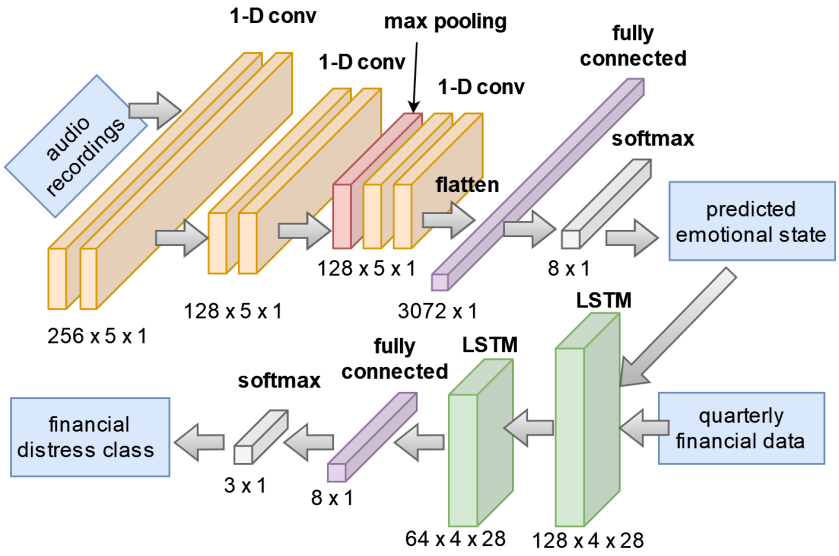

Fig. 3. Deep learning architecture for predicting financial distress.

In the next step, the output emotional state data from the SER module (8 emotional features) were merged with the financial data (20 features) collected from corporate financial statements to produce inputs for two LSTM layers with 128 and 64 neurons, respectively. To make the 1-year-ahead predictions of financial distress, four time steps (quarters) were used. Again, one fully connected dense layer was used to consolidate the output for the forecasted financial distress class (three classes were considered, namely, the safe zone, grey zone, and distress zone). Unlike the CNN architecture, we experimented with different LSTM architectures (one or two LSTM layers with 2^3, 2^4 to 2^9 neurons and one or two dense layers with 2^3, 2^4 to 2^8 neurons) to achieve the best classification performance. The learning configuration for the Adam optimizer was as follows: the learning rate of 0.0001, 100 epochs, and cross-entropy loss used as the fitness function. To conduct the experiments with the proposed deep learning architecture, we used the Keras library on a Jetson AGX Xavier Developer Kit equipped with 512-core Volta GPU with Tensor Cores and 32GB memory.

6 Results

Using the CNN-based SER system, we were able to achieve 69.8% accuracy on the RAVDESS dataset (using five-fold cross-validation), a result close to that achieved in [19]. The trained CNN-based SER system allowed us to assign emotional state labels to the audio recordings. A calm emotional state prevailed in 67% of the earnings conference calls, happiness followed with 10%, whereas emotional states of surprise or disgust were rarely found ($< 1\%$).

In the next step, we used two separate sets of experiments to predict the financial distress of companies. First, we tested the proposed model on a 1-year-ahead prediction of the three classes of financial distress (safe, grey, or distress). In the second set of experiments, we applied the model to the 1-year-ahead prediction of the trend of financial distress (upward or downward). To evaluate the contribution of the proposed CNN+LSTM prediction model, we compared the prediction results with a baseline model that did not take into account the emotional states of managers during their earnings conference calls. This baseline model is hereafter referred to as the LSTM model. To validate the model, we used several existing models that had been used in previous studies to predict financial distress. We used the values of the above 20 financial ratios 1 year in advance as input variables for all the models compared. Specifically, we used the following models for comparison:

- SMOTE+ADASVM [25] (the combination of the synthetic minority over-sampling technique (SMOTE) with the AdaBoost SVM (ADASVM) ensemble). Since the learning parameters are not presented in the original study, we tested both the linear and polynomial kernel functions for SVM base learners with different values of the regularization parameter $C = \{2^{-1}, 2^0, 2^1, \ldots, 2^5\}$.
- XGBoost [26]. As in [26], the learning rate was set to 0.1, the maximum tree depth for base learners was set to 10, and the subsample ratio was set to 0.7.
- Multilayer Perceptron (MLP) [27] with one hidden layer of 20 ReLUs, trained using the Adam optimizer with the learning rate of 0.001, the maximum number of epochs of 200, and the L2 penalty parameter of 0.0001.
- CUS+GBDT [28] (clustering-based under-sampling (CUS) combined with the gradient boosting decision tree (GBDT)). The number of clusters for CUS was set to 3, 100 estimators were used for GBDT, and the maximum depth of the individual regression estimators was 3.
- Stacking SVM [29]. In agreement with [29], the SVM with linear kernel functions was used to construct the base and meta classifiers. Again, different values of the regularization parameter C were examined.

The Imbalanced-learn library and Scikit-learn library were used for the experiments with the compared methods. To evaluate the performance of the models on the testing data, we used three standard measures of classification performance, that is accuracy (Acc), area under the receiver operating characteristic curve (AUC), and F1 measure (the weighted harmonic mean of precision and recall).

The results in Table 2 show that the existing prediction models provided a high accuracy in predicting financial distress 1 year ahead. Among these models, the best results were obtained by the XGBoost method, despite the fact that, unlike the SMOTE+ADASVM and CUS+GBDT methods, it does not address the problem of data imbalance in the financial distress classes. The high accuracy of our baseline model implied the benefit of a larger window size (time steps) in the proposed LSTM model. We were also able to slightly improve the accuracy by including managerial emotional states, suggesting that while emotional features may be valuable for predicting financial distress, financial ratios are crucial for this prediction. Moreover, the high AUC value indicates that the model performed well on all classes of financial distress. Finally, a balanced performance between precision (0.944 for class and 0.733 for trend prediction) and recall (0.943 and 0.733) was achieved for both prediction tasks.

Table 2. Results of 1-year-ahead financial distress prediction.

	Class prediction			Trend prediction		
Prediction model	Acc	AUC	F1 measure	Acc	AUC	F1 measure
SMOTE+ADASVM	92.04	0.967	0.920	60.30	0.653	0.603
XGBoost	94.15	0.988	0.942	70.02	0.766	0.700
MLP	92.59	0.975	0.926	60.55	0.634	0.605
CUS+GBDT	93.35	0.982	0.934	66.06	0.710	0.660
Stacking SVM	91.96	0.925	0.919	61.89	0.618	0.617
Our baseline LSTM model	94.27	0.990	0.943	72.80	0.789	0.728
Our CNN+LSTM model	**94.36**	**0.991**	**0.944**	**73.26**	**0.801**	**0.733**

Note: the best classification results are in bold.

The inclusion of emotional states had an even greater effect on increasing the classification accuracy when predicting the trend of financial distress. The results for trend prediction are generally consistent with those for financial distress class prediction. However, trend prediction seems to be a more complex task than predicting classes of financial distress. The results again confirmed the validity of the established model for predicting financial distress compared with existing models based on financial ratios only.

To better understand the effect of emotional states on prediction results, we used SHAP (SHapley Additive exPlanations) values, a game theoretic approach used to explain the output of deep learning models [30]. Among the greatest advantages of SHAP values is that they offer both global explainability (the overall decision structure of the model) and local explainability (how a decision is made in the case of individual samples). Here we focused on the global explainability of the model to demonstrate the impact of individual emotional states on the prediction of financial distress. We used the SHAP library to produce the SHAP values. In Fig. 4, we show the SHAP values of the eight emotional states compared with the most important financial indicator (the Z-score 1 year ago). The results show that the emotions of happiness and sadness were the most impor-

tant emotions for predicting the financial distress class, while the emotions of calm and anger were the most important emotions for predicting the financial distress trend. The emotion of happiness indicated a good financial situation, and the emotion of sadness indicated financial difficulty. For the trend prediction, the effect of emotions was not so clear. Nevertheless, the calm and happy emotions were indicative of an improved financial condition, whereas the angry and sad emotions were more likely to imply deterioration.

Fig. 4. SHAP values illustrating how emotional states contribute to the prediction of financial distress (class prediction on the left, trend prediction on the right).

7 Conclusion

The present study was designed to determine the effect of emotions in earnings conference calls on the prediction of corporate financial distress. The results of this study imply that state-of-the-art SER systems provide valuable information for financial distress prediction models. It was shown that by incorporating emotional states into the prediction model, the prediction performance can be improved within an appropriate deep learning-based architecture. Consistent with previous literature focused on sentiment analysis in text-based managerial communications [3–5], we observed that positive emotions (e.g., happiness) in audio recordings of earnings conference calls suggested a good financial situation of a company, whereas negative emotions (e.g., sadness and anger) indicated financial difficulties.

The main limitation of this study is that only one overall emotion was considered for the entire audio recording. The emotional states of speakers may change during a conference call, for example, when different topics are discussed. Therefore, in future research, we plan to analyze the content of the transcripts of the

earnings conference calls to be able to determine the topic being discussed. Incorporating emotions from the text of transcripts into a multi-modal prediction model is another interesting direction. In a future extension, fuzzy sets can also be used to represent the Z-score, and the prediction horizon can be extended to better account for dynamic business environment.

References

1. Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., Bilal, M.: Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications **94**, 164–184 (2018)
2. Hajek, P., Olej, V.: Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: Int. Conf. on Engineering Applications of Neural Networks, Springer, Berlin, pp. 1–10 (2013)
3. Hajek, P., Olej, V., Myskova, R.: Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. Technological and Economic Development of Economy **20**(4), 721–738 (2014)
4. Mai, F., Tian, S., Lee, C., Ma, L.: Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research **274**(2), 743–758 (2019)
5. Nguyen, B. H., Huynh, V. N.: Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. Journal of the Operational Research Society, 1–20 (2022)
6. Price, S. M., Doran, J. S., Peterson, D. R., Bliss, B. A.: Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance **36**(4), 992–1011 (2012)
7. Myskova, R., Hajek, P.: Mining risk-related sentiment in corporate annual reports and its effect on financial performance. Technological and Economic Development of Economy **26**(6), 1422–1443 (2020)
8. Mayew, W. J., Venkatachalam, M.: The power of voice: Managerial affective states and future firm performance. The Journal of Finance **67**(1), 1–43 (2012).
9. Price, S. M., Seiler, M. J., Shen, J.: Do investors infer vocal cues from CEOs during quarterly REIT conference calls?. The Journal of Real Estate Finance and Economics **54**(4), 515–557 (2017)
10. Hobson, J. L., Mayew, W. J., Venkatachalam, M.: Analyzing speech to detect financial misreporting. Journal of Accounting Research **50**(2), 349–392 (2012)
11. Yang, L., Ng, T. L. J., Smyth, B., Dong, R.: Html: Hierarchical transformer-based multi-task learning for volatility prediction. In: Proc. of The Web Conference 2020, pp. 441–451 (2020)
12. Sawhney, R., Khanna, P., Aggarwal, A., Jain, T., Mathur, P., Shah, R.: VolTAGE: volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 8001–8013 (2020)
13. Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., Collins, L. M.: Financial fraud detection using vocal, linguistic and financial cues. Decision Support Systems **74**, 78–87 (2015)
14. Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., Byrd, M., Spitzley, L.: Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. Journal of Language and Social Psychology **35**(2), 123–157 (2016)

15. Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., Zimmermann, R.: Multi-modal multi-task financial risk forecasting. In: Proc. of the 28th ACM Int. Conf. on Multimedia, pp. 456–465 (2020)
16. Sawhney, R., Aggarwal, A., Khanna, P., Mathur, P., Jain, T., Shah, R. R.: Risk forecasting from earnings calls acoustics and network correlations. In: INTERSPEECH, pp. 2307–2311 (2020)
17. Sawhney, R., Aggarwal, A., Shah, R.: An empirical investigation of bias in the multimodal analysis of financial earnings calls. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3751–3757 (2021)
18. Cao, S., Jiang, W., Yang, B., Zhang, A. L.: How to talk when a machine is listening: Corporate disclosure in the age of AI. National Bureau of Economic Research, no. w27950 (2020)
19. Issa, D., Demirci, M. F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control **59**, 101894 (2020)
20. Livingstone, S. R., Russo, F. A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS One 13(5), e0196391 (2018)
21. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: Librosa: Audio and music signal analysis in python. In: Proc. of the 14th Python in Science Conf., vol. 8, pp. 18–25 (2015)
22. Hajek, P., Michalak, K.: Feature selection in corporate credit rating prediction. Knowledge-Based Systems **51**, 72–84 (2013)
23. Son, H., Hyun, C., Phan, D., Hwang, H. J.: Data analytic approach for bankruptcy prediction. Expert Systems with Applications **138**, 112816 (2019)
24. Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., Suvas, A.: Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. Journal of International Financial Management & Accounting **28**(2), 131–171 (2017)
25. Sun, J., Li, H., Fujita, H., Fu, B., Ai, W.: Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion **54**, 128–144 (2020)
26. Huang, Y. P., Yen, M. F.: A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing **83**, 105663 (2019).
27. Alaminos, D., Fernández, M. Á.: Why do football clubs fail financially? A financial distress prediction model for European professional football industry. PloS One **14**(12), e0225989 (2019)
28. Du, X., Li, W., Ruan, S., Li, L.: CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. Applied Soft Computing **97**, 106758 (2020)
29. Liang, D., Tsai, C. F., Lu, H. Y. R., Chang, L. S. Combining corporate governance indicators with stacking ensembles for financial distress prediction. Journal of Business Research **120**, 137–146 (2020)
30. Lundberg, S. M., Lee, S. I.: A unified approach to interpreting model predictions. In: Proc. of the 31st Int. Conf. on Neural Information Processing Systems, pp. 4768–4777 (2017)