



# Speech emotion recognition and text sentiment analysis for financial distress prediction

Petr Hajek<sup>1</sup> · Michal Munk<sup>1,2</sup>

Received: 28 December 2022 / Accepted: 7 March 2023 / Published online: 20 March 2023  
© The Author(s) 2023

## Abstract

In recent years, there has been an increasing interest in text sentiment analysis and speech emotion recognition in finance due to their potential to capture the intentions and opinions of corporate stakeholders, such as managers and investors. A considerable performance improvement in forecasting company financial performance was achieved by taking textual sentiment into account. However, far too little attention has been paid to managerial emotional states and their potential contribution to financial distress prediction. This study seeks to address this problem by proposing a deep learning architecture that uniquely combines managerial emotional states extracted using speech emotion recognition with FinBERT-based sentiment analysis of earnings conference call transcripts. Thus, the obtained information is fused with traditional financial indicators to achieve a more accurate prediction of financial distress. The proposed model is validated using 1278 earnings conference calls of the 40 largest US companies. The findings of this study provide evidence on the essential role of managerial emotions in predicting financial distress, even when compared with sentiment indicators obtained from text. The experimental results also demonstrate the high accuracy of the proposed model compared with state-of-the-art prediction models.

**Keywords** Sentiment analysis · Speech emotion recognition · FinBERT · Deep learning · Financial distress · Earnings conference calls

## 1 Introduction

Models for predicting financial distress are increasingly considered among the most valuable financial models as they are capable of providing an early warning signal to stakeholders of imminent corporate failure. Past financial crises incurred considerable financial losses, heightening

the urgency to alleviate the information asymmetry between firm managers and the remaining stakeholders by providing them with powerful financial distress prediction models.

The overall objective of financial distress prediction is to foresee whether a company will run into financial difficulties, using statistical or machine learning models based on quantitative or qualitative financial indicators currently available. Most previous models were designed with the aim of accurate detection one year before failure [3]. Many traditional models fail for longer prediction horizons [5]. However, some models are able to predict financial distress earlier, typically three to five years before failure, with relatively high accuracy [12, 54]. Financial distress prediction plays a fundamental role in a number of decision-making problems faced by corporate stakeholders, including corporate financial managers' decisions, investors'

---

✉ Petr Hajek  
petr.hajek@upce.cz

Michal Munk  
mmunk@ukf.sk

<sup>1</sup> Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, 532 10 Pardubice, Czech Republic

<sup>2</sup> Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia

investment decisions, lenders' credit decisions and banks' assessment of clients' creditworthiness [56].

Most financial distress prediction models have so far tended to concentrate on financial indicators derived from corporate financial statements [3]. Furthermore, some studies focusing on SMEs in particular have considered external drivers unrelated to managerial skills [17, 19], such as psychological factors (e.g. psychological distress of owners), natural disasters, regulatory inefficiencies or labour market turbulence. However, with the increasing availability of textual data relating to companies, there is an emerging opportunity to examine whether indicators derived from these textual documents can improve the prediction of financial distress. Further insights regarding the attitudes of managers and investors have been found to be significant indicators of firms' financial performance. Prior relevant research in this respect has concentrated on extracting textual sentiment from managerial discussion included in corporate disclosures [22, 24, 38, 44]. However, linguistic indicators derived from annual reports suffer from low frequency and, therefore, outdatedness, which limits them in dynamical financial distress prediction. In addition to annual reports, other sources of managerial communication are available. In particular, earnings conference calls provide a more flexible source of information with more forward-looking narratives, as they also include communication between executives and analysts [46]. Indeed, the linguistic tone of conference call transcripts proved to be a prominent indicator of abnormal returns [15]. Furthermore, previous studies revealed that the effect of emotional linguistic information may outweigh the value of factual accounting information as disclosed by executives, thereby indicating underlying company financial risks [43]. Evidence from empirical studies also suggests that nonverbal managerial communication, when coupled with financial indicators, is incrementally useful in financial performance prediction [40, 47]. This can be attributed to the cognitive conflict (dissonance) between actual financial performance and managerial emotional state, suggesting potential financial difficulties.

Motivated by the above, this study asserts that sentiments and emotions derived from earnings conference calls can provide important indicators of financial distress. To this end, this study aims to incorporate linguistic indicators and vocal cues extracted from earnings conference calls in models predicting corporate financial distress. To address this issue, a deep learning model is used that incorporates: (1) a set of diverse dictionary-based sentiment and emotion indicators and a domain-specific FinBERT model to represent conference call transcripts; (2) a convolutional neural network (CNN)-based speech emotion recognition model proposed by [29] to capture the managerial

emotional states in conference call recordings; and (3) a long short-term memory (LSTM) neural network leveraged by the linguistic indicators, emotional states and financial indicators for dynamical financial distress prediction. By fusing verbal and nonverbal managerial communication obtained from the transcripts and audio recordings of earnings conference calls, respectively, and financial ratios from corporate financial statements, we believe that the fused model will dominate the traditional prediction models using only financial indicators. Furthermore, by comparing the predictive performance of verbal and nonverbal attributes, this study aims to enhance our understanding of the role of text sentiment and speech emotion in financial distress prediction.

In summary, the contributions of this paper are fourfold:

- A novel dynamical financial distress prediction model is proposed by using LSTM neural networks. Unlike the existing stationary financial distress prediction models based on linguistic attributes, the proposed model utilises a more frequent source of textual data represented by earnings conference calls.
- For the first time, a rich word representation combining dictionary-based emotion features and domain-specific transformer-based sentiment model is used for financial distress prediction. To this end, a high lexical coverage is ensured by using a wide range of emotion indicators, as well as that based on a pre-trained FinBERT model.
- To the best of our knowledge, this is the first study using speech emotion recognition for financial distress prediction. Existing research based on vocal features was limited to stock price prediction and financial fraud prediction.
- To show the predictive power of verbal and nonverbal components of the model, several sets of experiments are performed to predict the financial distress of the largest US companies. As far as we know, this is the first time that a combination of verbal and nonverbal features are used for financial distress prediction.

This paper is a substantially extended version of [20]. Note that the earlier version was limited to speech emotions, while neglecting the sentiment and emotion information provided in conference call transcripts. Furthermore, an improved deep learning architecture for financial distress prediction is proposed in the current study and the effect of different approaches to sentiment analysis is investigated. This extension also presents an in-depth comparative analysis with existing financial distress prediction models based on textual data.

The remainder of this paper is organised as follows. Section 2 provides the literature review on using linguistic indicators and vocal cues in the field of finance. Section 3 presents the components of the proposed conceptual

framework for financial distress prediction. Section 4 describes the data used. Section 5 describes the setting of the proposed deep learning architecture and presents the experimental evidence for the excellent performance of the proposed model, especially compared with existing prediction model using only financial indicators. Section 6 concludes the paper and suggests future research directions.

## 2 Related work

The literature review provided in this section focuses on two main lines of research in the field of sentiment and emotion analysis in finance, namely the use of (1) linguistic sentiment indicators obtained through text analysis and (2) emotional states extracted through speech emotion recognition.

### 2.1 Financial distress prediction using linguistic indicators

Early works in the area of financial distress prediction focused primarily on quantitative financial indicators extracted from financial statements [3, 56]. The study of Cecchini et al. [11] has made a significant contribution to the field by examining, for the first time, the effect of information obtained from financial texts on the accuracy of predicting financial distress. More precisely, terms with a high frequency of occurrence and high discriminating power were identified in MD&A (management discussion and analysis) sections of corporate annual reports, thus complementing the quantitative financial indicators. As a result, the improvement in prediction accuracy was an impressive 17.1% over the baseline model consisting of the Altman model's financial ratios. Notably, even using text alone, an accuracy of 80.0% was achieved, outperforming the often-used Altman model. Limitations of the methodology used by Cecchini et al. [11] stem from the bag-of-words approach, leading to sparse and high-dimensional document representation. Moreover, the reliability of the discriminatory power of the words contained in the annual reports of companies with financial distress is limited by the small number of such companies.

To overcome the problems of high dimensionality and sparsity of the document representation using the bag-of-words approach, dictionary-based approaches have been preferred in subsequent studies [22, 44]. These, among other things, facilitate the interpretability of models by reducing the information from the bag-of-words model into typically two linguistic indicators, i.e. positive and negative sentiment. However, general dictionaries, such as the Harvard's General Inquirer, can be context-sensitive, so

dictionaries developed for finance have proved to be more relevant. In their widely acclaimed work, Loughran and McDonald [36] proposed a set of dictionaries for the finance domain, including those for positive and negative financial sentiment. This has allowed other researchers to more accurately estimate sentiment in annual reports (especially managerial sentiment in the MD&A sections) and to achieve more accurate predictions of credit ratings [22], overall corporate financial performance [24] and corporate bankruptcy [23]. These studies also offered sensitivity analyses comparing general and financial dictionaries, concluding that the latter are more indicative predictors of financial distress.

To achieve a richer linguistic representation of annual reports, Wang et al. [62] combined the dictionary-based approach with a bag-of-words model based on term frequency. Specifically, the model using financial features performed best with AUC=0.912, followed by the bag-of-words model (AUC=84.5) while the financial dictionary-based model performed the worst among these individual models with AUC=0.681, suggesting that dictionary-based linguistic indicators, while easy to interpret, do not provide sufficient predictive power on their own. However, when the features of these approaches were fused, the overall performance of financial distress prediction could be significantly increased to AUC=0.961, but this was mainly due to the merit of the bag-of-words linguistic features.

In addition to the above problems of bag-of-words models with high data dimensionality and sparsity, the inadequate consideration of the context of financial documents should be pointed out. To overcome these problems, low-dimensional dense feature representations based on word embeddings were used in more recent studies. Specifically, the word2vec model was used by Mai et al. [38] to effectively capture word meaning and word context of annual reports. By simply averaging the word embeddings in each document and adding two fully connected layers, a relatively accurate financial distress prediction model was produced that outperformed more complex models using convolutional neural networks. However, Matin et al. [39] and Huang et al. [27] reached the conclusion that word embeddings of auditors' reports are more informative than those based on managerial statements in annual reports. The main drawback of the word2vec model is that it does not support explicit sentence representation. In contrast, the BERT language model generates contextual sentence and word representation. Li et al. [32] exploited the advantages of the BERT model when constructing a financial domain sentiment dictionary, which outperformed the traditional dictionary of Loughran and McDonald [36]. Most recently, Jiang et al. [30] have showed that semantic features extracted from the BERT model have stronger predictive power than topic and sentiment features.

A list of related studies is presented in Table 1, showing the method used, the data and the resulting classification accuracy.

Overall, earlier research reported that the textual information presented in the annual report is qualitatively complementary to the data obtained from financial statements and can therefore be of great importance to stakeholders. The purpose of the annual report is to provide additional information on the company's financial situation and to outline the future financial development of the company. In addition, management priorities and financial risks are discussed. However, the above studies using annual reports as text source have several major shortcomings. The main one is the low frequency of annual reports. Hence, linguistic indicators extracted from annual reports are not ideal for time series prediction. As a result, the above studies represent only stationary financial distress prediction models, neglecting the impact of concept drift in financial distress prediction. In fact, the area of dynamic financial distress prediction is attracting considerable interest due to its capacity to take into account the change in the distribution of financial data over time [1, 57, 58]. To capture the dynamics of the change, we used LSTM neural networks in this study.

The low frequency of disclosure of annual reports is linked to the lack of timeliness of the data contained therein. To provide the financial distress prediction model with more up-to-date dynamic data with higher frequency, we use earnings conference calls as the text source in the current study. Similar to annual reports, conference calls

capture insider perspectives and views, reducing the potential information asymmetry between investors and managers. In addition, earnings conference calls are a more flexible source of information than annual reports, and the conversation between executives and analysts regarding corporate performance is a potentially rich source of information [46]. Furthermore, earnings conference calls typically contain more forward-looking narratives [8], and therefore, the linguistic features in these transcripts have more potential to predict future corporate financial performance.

Although BERT represents a state-of-the-art language model, sentiment analysis is a heavily domain-dependent task. Therefore, to leverage the large amount of financial text, including corporate reports, earnings call transcripts and analysts' reports, FinBERT was pre-trained as a domain-specific BERT model. Results reported by Araci [6] and Yang et al. [64] suggest that FinBERT outperforms the generic domain BERT model on financial sentiment analysis tasks. Another limitation of previous studies is their tendency to focus on polarity-oriented sentiment, using lexicons of positive and negative words. Not only do these lexicons usually lack the intensity of the sentiment level, but narrowing the scope of linguistic analysis to only positive and negative sentiment is inadequate in the case of earnings conference calls, as management typically displays a much broader range of emotions during discussions with analysts, such as anger, surprise, fear, among others [46]. Therefore, to provide the financial distress prediction model with a state-of-the-art finance-specific

**Table 1** Financial distress prediction using textual data

Study	Method	Features	Text source	Performance
Cecchini et al. [11]	SVM	Bag-of-words	Annual reports	Acc = 0.839
Hajek and Olej [22]	SVM	L&M	Annual reports	Acc = 0.924, F1 = 0.924
Hajek et al. [24]	SVM	L&M	Annual reports	Acc = 0.838, F1 = 0.834, AUC = 0.891
Hajek and Olej [23]	NBTrees	L&M	Annual reports	Acc = 0.951, F1 = 0.948, AUC = 0.989
Wang et al. [62]	RanSub	HowNet sentiment	Annual reports	Acc = 0.964, AUC = 0.952
Matin et al. [39]	CNN+LSTM	word2vec	Annual reports	AUC = 0.844
Mai et al. [38]	CNN	word2vec	Annual reports	Acc = 0.712, AUC = 0.856
Tang et al. [60]	RNN	pos./neg. sentiment	Annual reports	Acc = 0.939, F1 = 0.934, AUC = 0.936
Wang et al. [62]	RanSub	L&M + bag-of-words	Annual reports	AUC = 0.961
Li et al. [32]	DNN	BERT	Annual reports	Acc = 0.901, F1 = 0.896
Li et al. [32]	DNN	L&M	Annual reports	Acc = 0.857, F1 = 0.852
Huang et al. [27]	XGBoost	word2vec	Annual reports	Acc = 0.921, F1 = 0.494, AUC = 0.912
Zhao et al. [66]	CatBoost	Sentiment score	Annual reports	AUC = 0.976
Jiang et al. [30]	RF	BERT	Annual reports	AUC = 0.936
This study	LSTM	Emotions+FinBERT sentiment	Earn. conf. calls	

CNN Convolutional neural network, L&M sentiment indicators from the L&M dictionary [36], LSTM long short-term memory, RanSub random subspace, RF random forest, RNN recurrent neural network, SVM support vector machine

language model and fine-grained sentiment indicators, the combination of FinBERT sentiment and emotion features allowed us to constitute a rich linguistic representation of earnings conference calls.

## 2.2 Vocal features in finance

A considerable amount of the literature has been published on the role of vocal cues in finance in the past decade. Earlier research has shown that financial prediction models can benefit from vocal features derived from earnings conference calls, see Table 2. Considering the vocal features used, earlier research tended to apply two fundamental voice analysis tools: LVA (layered voice analysis) [26, 40] and Praat [50, 63].

LVA allowed the researchers to derive affective state levels, including the cognitive, thinking, emotional and stress level. The cognitive level represents cognitive dissonance, while the emotional level stands for the level of excitement. The thinking level denotes mental efforts, and physical arousal is estimated using the stress level. In particular, several studies using LVA reported that extremely high emotional levels are significant indicators of abnormal financial states. Mayew and Venkatachalam [40] and Price et al. [47] reported that excessive affective states

indicate future cumulative abnormal returns of stocks, while vocal dissonance can be used to identify financial fraud [26].

The Praat voice analysis tools offer a different perspective than LVA, providing users to quantify a variety of acoustic features, such as voice intensity, pitch, shimmer, jitter and excitation. Throckmorton et al. [61] revealed that combining linguistic and vocal features outperformed financial indicators in fraud detection. Similarly, testimonies concealing deception allegedly had a higher pitch and lower voice quality than legitimate testimonies [9]. Vocal cues are also reported to have a good predictive capacity for stock price volatility [63]. In a similar manner, a multi-modal prediction model also worked well in predicting stock price volatility Sawhney et al. [50], as well as the neural attentive alignment model capturing interdependencies between verbal and vocal modalities [49, 51]. The most recent work found that there is gender bias in multi-modal financial forecasts [52] and that managers modify the way they talk knowing that machines are listening [10].

Summarising the previous findings, vocal features have so far only been used in detecting financial fraud and predicting stock price volatility and abnormal stock return. In addition to the new application area of vocal cues, this

**Table 2** Summary of prior studies using vocal features in finance

Study	Method	Features	Prediction task
Hobson et al. [26]	LVA	Cognitive dissonance	Financial misreporting
Mayew and Venkatachalam[40]	LVA	Positive and negative affect	CAR
Throckmorton et al. [61]	Praat	Fundamental frequency, variations of amplitude maxima, small-scale perturbations, proportion of voiced speech, mean harmonics-to-noise ratio	Financial fraud
Burgoon et al. [9]	SPLCE	Vocal intensity, response latency, pitch slope, voice quality	Fraudulent utterances
Price et al. [47]	LVA	Emotional and cognitive activity level	CAR
Sawhney et al. [49]	Praat	26 vocal cues	Stock price volatility
Yang et al. [63]	Praat	27 vocal cues	Stock price volatility
Cao et al. [10]	pyAudioAnalysis	Emotion arousal and valence	AI readership
Sawhney et al. [50]	Praat	26 vocal cues	Stock price volatility
Sawhney et al. [51]	Praat	26 vocal cues	Stock price
Sawhney et al. [52]	Praat	26 vocal cues	Stock price volatility
This study	CNN-based SER	8 emotional states based on 180 spectral features	Financial distress

CAR Cumulative abnormal return, LVA layered voice analysis, SER speech emotion recognition, SPLCE structured programming for linguistic cue extraction

study takes advantage of recent developments in the field of speech emotion recognition. Specifically, deep learning allows us to accurately identify the emotional states of managers based on a wide range of spectral features.

### 3 Model for financial distress prediction

The conceptual framework outlined in the current study for predicting financial distress is represented by a deep learning model combining three types of information: (1) managers’ emotions obtained from their speech during conference calls; (2) sentiment obtained from the transcription of the conference calls; and (3) conventional financial indicators from corporate financial statements. This section details the model.

#### 3.1 Deep learning model for financial distress prediction

The proposed financial distress prediction model is presented in Fig. 1. The deep learning architecture comprises the pre-trained CNN model that classifies the audio recordings from earnings conference calls into 8 emotional states using a wide range of spectral features. The CNN model for speech emotion recognition is derived from the model proposed by Issa et al. [29] and adapted for eight emotional states. Specifically, the first two one-dimensional convolutional layers contained 256 and 128 filters, respectively. As depicted in Fig. 1, the kernel size was 5 and stride was 1. The max-pooling layer followed using a window size of 8. Next, another one-dimensional convolutional layer was used, and flattening resulted in a fully connected layer followed by a dropout layer to prevent over-fitting (dropout rate was 0.2). Finally, a softmax layer

was used to produce the predicted emotional states. The Adam optimizer was used to train the CNN model using the cross-entropy loss function and a learning rate of 0.0001. Note that to obtain the managerial emotional states, the CNN model was trained on a labelled benchmark dataset using fivefold cross-validation. Then, the trained model was used to classify the unlabelled managers’ audio recordings.

Next, the emotional state features were fused with the 20 financial features from quarterly financial statements and FinBERT-based sentiment and emotional states extracted from conference call transcripts. Thus, inputs for LSTM layers were produced. The LSTM model was employed to capture high-level temporal features from sequential quarterly corporate data [14, 34]. In contrast to the CNN model, whose structure was taken from a previous study [29], different LSTM structures were examined to achieve the best prediction performance. More precisely, we tested the LSTM model with one or two layers comprising  $2^3$ ,  $2^4$  to  $2^9$  units and one or two fully connected layers with  $2^3$ ,  $2^4$  to  $2^8$  units. The LSTM model was trained to predict the financial distress class (safe/grey/distress) one year ahead. Alternatively, the softmax layer with two units was used to predict the financial performance trend (upward/downward). After testing different model structures, the final LSTM model consisted of two layers with  $2^7$  and  $2^6$  units and one fully connected layer with  $2^3$  units, respectively. The Adam optimizer was used with the cross-entropy loss function, the learning rate was 0.0001, and the number of epochs was 100. Again, this setting of the hyperparameters was the result of experimentation with their various values, namely the learning rate with values of 0.01, 0.001 and 0.0001, and the number of epochs of 10, 100 and 500. For the experiments with the proposed deep learning model, the Keras library was used running on a Jetson AGX Xavier (512-core Volta GPU, 32GB memory).

#### 3.2 Speech emotion recognition

As highlighted above, prior research has only looked at the vocal utterances of managers, not directly gauging the emotions managers express in earnings conference calls. To tackle this issue, we first used a CNN model for speech emotion recognition. It is worth noting that various models have been developed and introduced to conduct speech emotion recognition [2]. It was decided that the best model to adopt for this investigation was the deep CNN model proposed by [29] because it delivered state-of-the-art performance for diverse speech emotion datasets, including the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset [35]. This dataset was chosen to pre-train the CNN model in this study because it

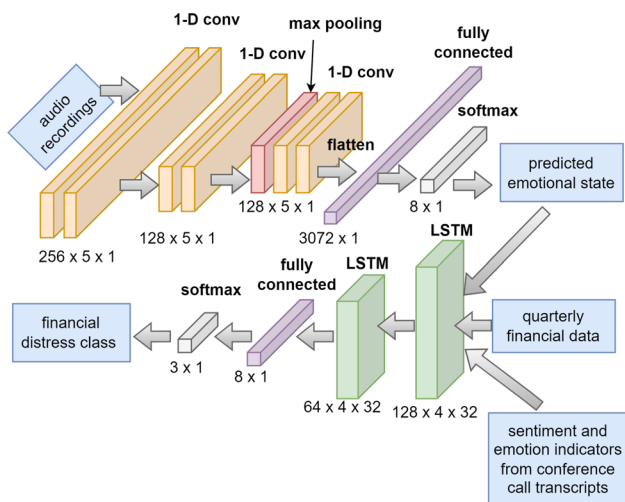


Fig. 1 Deep learning model for financial distress prediction

is considered a validated and genuine benchmark dataset in the field of speech emotion recognition [16, 18, 53].

The RAVDESS dataset includes 1,440 audio recordings. When building the dataset, different levels of emotional intensity were considered and each recording was evaluated in terms of emotional validity, genuineness and intensity to provide researchers with a gold standard for speech emotion recognition [35]. The dataset comprises eight emotional states: calm, neutral, surprised, happy, angry, sad, disgust and fearful. It is worth noting that the dataset is almost balanced in terms of the frequency of individual emotions. (Only the neutral emotional state includes 96 recordings, and other emotional states are equally represented by 192 audio recordings.)

To obtain features for speech emotion recognition, the Librosa audio library [41] was utilised. More precisely, 180 spectral features were extracted, divided into three distinct feature categories: (1) 40 features representing mel-frequency cepstral coefficients, (2) 12 chromagram features and (3) 128 mel-scaled spectrogram features. This set of features was used to mimic the way humans receive sound frequencies [29]. Mel-frequency cepstral coefficients produce a mel-frequency cepstrum, thus representing the short-term sound power spectrum. Short-time Fourier transform is used to obtain chromagrams, constituting harmony and pitch classes [41]. The mel-scaled spectrogram features are obtained using a nonlinear transformation of the spectrogram to simulate how relative frequencies (emotions) are perceived by a human. Altogether, our aim was to obtain a rich audio recording representation for the CNN model.

After pre-training the CNN model on the RAVDESS dataset, we fed the CNN model with the unlabelled audio recordings of earnings conference calls to get the labels of managers' emotional states.

### 3.3 Text sentiment analysis

As noted above, this study exploited the combination of domain-specific sentiment analysis and emotional features to produce a rich linguistic representation of conference call transcripts that adequately cover a broad range of managerial sentiments and emotions.

For the domain-specific sentiment analysis, we chose the pre-trained FinBERT due to its remarkable performance in sentiment analysis of various financial texts [6]. The FinBERT model for sentiment analysis comprised two steps. First, the authors used a large financial corpus of Reuters TRC2-financial with almost 30 M words to pre-train the BERT model [6]. Second, the model was fine-tuned for sentiment analysis using the Financial PhraseBank dataset, an annotated set of 4,845 sentences from financial news. Specifically, the annotators were requested to assign labels

based on how they thought the information in the financial news would be reflected in the stock price of the company mentioned. The fine-tuned pre-trained FinBERT model<sup>1</sup> was used to label each sentence in conference call transcripts. That is, we first performed sentence segmentation (using the Stanza tools [48]), and then, the intensity of conference call sentiment was calculated as the sentiment score averaged over all sentences.

It was considered that text emotional features would usefully supplement and extend both the speech emotions and the sentiment analysis, as reported in previous business research [25, 31]. To obtain the text emotional features, the following emotional states were considered in agreement with the Plutchik wheel of emotions [42]: anticipation, surprise, joy, trust, sadness, disgust, fear and anger. The text emotional features were calculated as the sum of emotional words matching the respective categories in pre-defined lexicons. The NRC emotion lexicon<sup>2</sup> was used for analysing emotions in conference call transcripts.

### 3.4 Financial features

The obtained audio and text features were matched with traditional financial features. To calculate the financial features, corporate financial statement data were used. The publicly available Macrotrends database<sup>3</sup> was used as the source of the financial data. Consistent with relevant studies on the prediction of financial distress [21, 55], the financial features comprised the following financial categories: (1) company size (sales, equity, total assets, cash flow); (2) activity ratios (inventory turnover, receivable turnover, asset turnover); (3) leverage ratios (total debt, equity to liabilities); (4) profitability ratios (gross margin, operating margin, return on equity, return on total assets, retained earnings to total assets); (5) liquidity ratios (free cash flow per share, operating cash flow per share, working capital to total assets, cash ratio, current ratio); and (6) global measure of financial performance (Altman's Z-score).

Company financial distress is generally regarded as a company's inability to comply with its payment obligations under the debt. Since the Altman's Z-score is the preferred model for the early detection of financial distress in the literature [5], we opted for this model to obtain the financial distress classes. Specifically, the model categorises companies into three zones: safe, grey and distress. Another advantage of this model is its validity for firms listed on US financial markets [5]. Furthermore, the model was shown to perform well on bankruptcy prediction tasks,

<sup>1</sup> <https://github.com/ProsusAI/finBERT>.

<sup>2</sup> <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>.

<sup>3</sup> [www.macrotrends.net](http://www.macrotrends.net).

accurately predicting bankruptcy up to 2 years ahead. The formula for the Z-score model for publicly tradeable companies is as follows:

$$Z\text{-score} = 1.2 \times \frac{WC}{TA} + 1.4 \times \frac{RE}{TA} + 3.3 \times \frac{NI}{TA} + 0.6 \times \frac{E}{TL} + 1.0 \times \frac{TS}{TA}, \quad (1)$$

where  $TA$  is total assets,  $WC$  is working capital,  $RE$  denotes retained earnings,  $NI$  is net income,  $E$  represents equity,  $TL$  denotes total liabilities and  $TS$  is total sales. A Z-score of 3 or higher denotes the safe class (healthy companies with a high survival probability), that of 1.80 to 2.99 indicates the grey zone (signalling some financial difficulties), and that of less than 1.80 denotes the distress class (high-risk companies with financial difficulties).

## 4 Data

The database of audio recordings comprised 1278 earnings conference calls gathered from the publicly available EarningsCast database<sup>4</sup> from Q1 2010 to Q3 2021 on a quarterly basis. The data sample included the 40 largest US companies listed on the NYSE (New York Stock Exchange) in terms of market capitalisation. In the pre-processing stage, the collected audio recordings were first converted to .wav files. Note that the audio recordings were supplied without any breakdown or identification of the speakers. That is why the whole record was labelled with one emotional state based on a set of 180 spectral features. To illustrate the features used for speech emotion recognition, Fig. 2 presents the spectral features obtained from the earnings conference call for Adobe. But before we could label these conference calls, it was necessary to train the speech emotion recognition model. To this end, the RAVDESS dataset was used to train the CNN model. As a result, 69.8% accuracy was achieved on the RAVDESS dataset using fivefold cross-validation. We adopted this testing scheme in keeping with the original study [29], yet we did not observe many differences in classification performance, even for other data split ratios and tenfold cross-validation. It is worth noting that this performance is close to that reported by Issa et al. [29]. We also examined the performance of other models that have previously shown promising accuracy on the RAVDESS dataset. More precisely, we obtained the following results using alternative approaches: 58.4% accuracy for bagged ensemble of SVMs [7], 59.8% for AdaBoost ensemble of SVMs [7] and 64.5% for deep residual networks [65]. Figure 3 shows that the

classification performance was reasonable across all the eight emotional states.

The trained CNN model was then used to label the conference call recordings with emotional states. As presented in Fig. 4, a calm emotional state dominated in the audio recordings (67%), followed by happiness (10%). In contrast, disgust and surprise rarely occurred (< 1%), see Fig. 4.

Regarding the text sentiment and emotional features, Table 3 shows that the overall sentiment of the conference call transcripts was more positive and that managers tended to use trustworthy and anticipatory tones while avoiding negative emotions, such as disgust and anger.

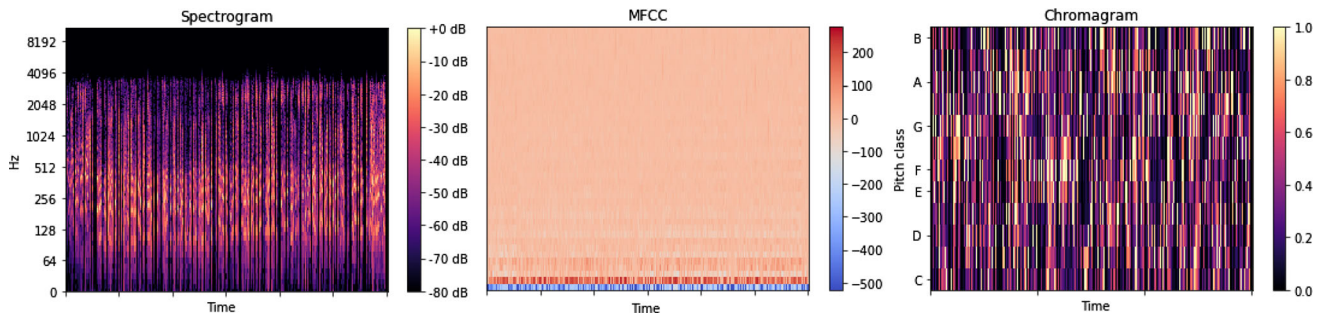
The overall financial performance of the companies is depicted in Fig. 5, indicating a deterioration in corporate financial performance from 2019. Indeed, the global economy entered into a challenging period of recession while recording the lowest growth of the decade due to decelerating domestic investment and trade disputes.

The companies were categorised into three financial distress classes according to the Z-score, with 59 companies belonging to the safe class (4.6%), 329 falling into the grey zone (25.7%) and 890 classified as distressed (69.7%), suggesting a problem of class imbalance. Interestingly, according to Altman et al. [5], the high percentage of firms in the distress class may be indicative of an impending financial crisis. The dataset was constructed in order to perform a 1-year-ahead prediction of financial distress. To validate the proposed model, an additional run of experiments was considered, categorising the companies into two classes, according to whether there was an increase or decrease in the overall financial performance (Z-score) in the next year. Regarding the trend of the financial performance, 615 companies were classified in the upward class, while 663 companies were put into the downward class.

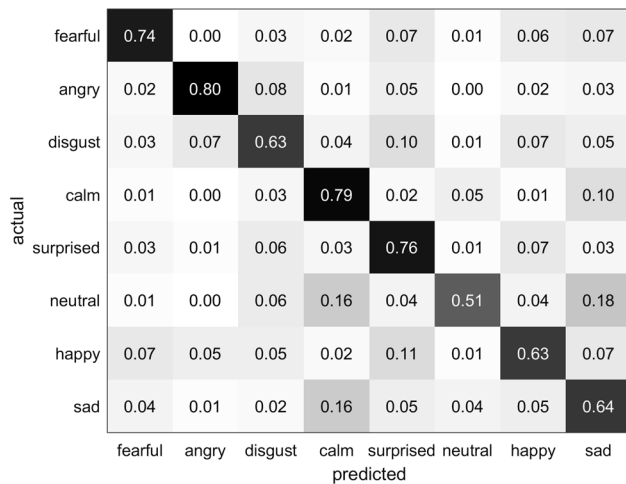
## 5 Model validation

To evaluate the effectiveness of the proposed model, two separate experimental runs were performed. In the first run, this study examined the model performance in a prediction of three financial distress classes (safe companies/grey zone companies/distressed companies). In the second run, the model was trained to forecast the upward or downward trend of financial distress. Both sets of experiments considered the 1-year-ahead prediction. Data partitioning took into account the temporal nature of the data, using data for 2010 to 2016 as training set and data for 2017 to 2020 as test set. To estimate the performance of financial distress prediction models, consistent with the previous literature [27, 32], a combination of three measures was considered in the current study. Specifically, accuracy (Acc), F1

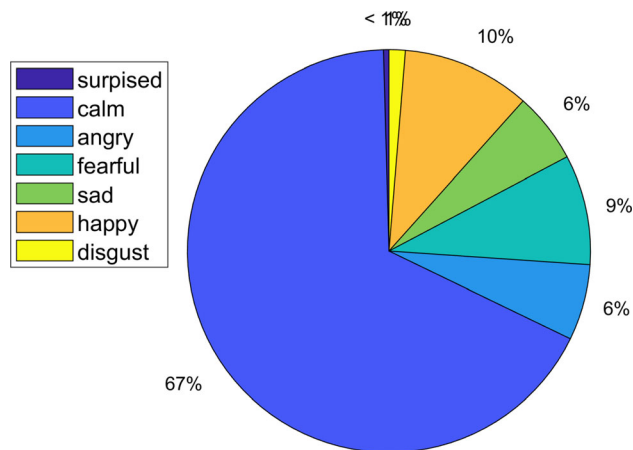
<sup>4</sup> <https://earningscast.com/>.



**Fig. 2** Example of spectral features (mel-scaled spectrogram, mel-frequency cepstral coefficients (MFCC) and chromagram) extracted from the earnings conference call of Adobe in Q4 2019



**Fig. 3** Confusion matrix of the CNN speech emotion recognition model for the RAVDESS dataset



**Fig. 4** Frequency of emotional states in the conference call dataset

measure and AUC (area under the receiver operating characteristic curve) were used to evaluate the classification performance of the model. The F1 measure is represented by the weighted harmonic mean of precision and sensitivity, where precision denotes the percentage of correctly identified distressed companies of all distressed

**Table 3** Descriptive statistics for linguistic features

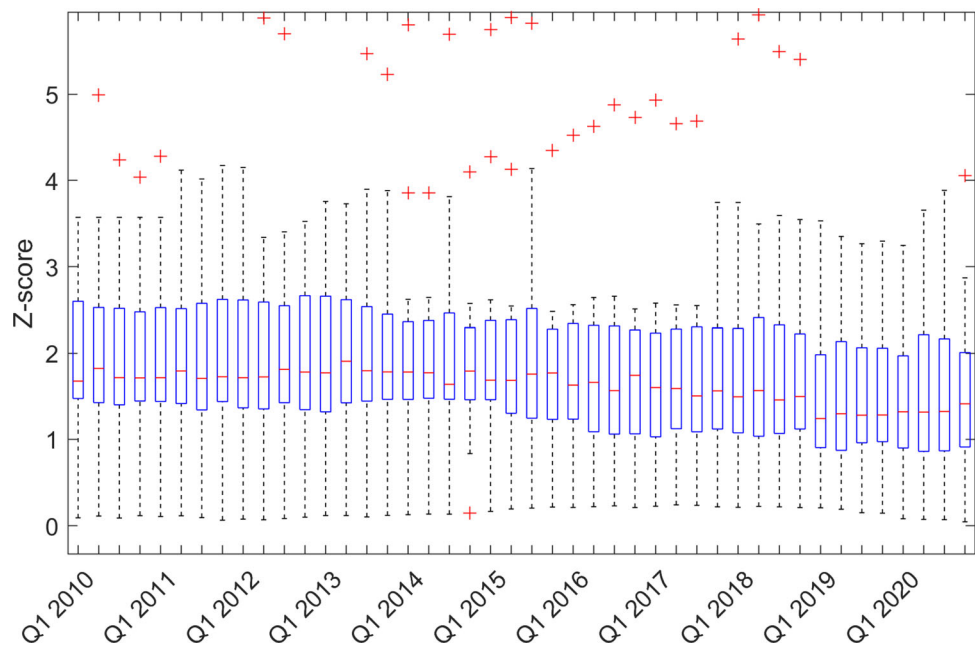
Linguistic feature	Mean	St. dev.
FinBERT sentiment	0.282	0.205
Anger	49.532	23.651
Anticipation	240.835	86.966
Disgust	29.338	20.584
Fear	65.153	33.028
Joy	138.731	52.575
Sadness	65.361	32.121
Surprise	66.647	26.967
Trust	314.945	111.575

The FinBERT sentiment was obtained as an average of all sentences in the conference call transcript, and the values of the remaining linguistic features represent the counts of emotional words in the transcript matching the NRC emotion lexicon

companies, and sensitivity indicates the percentage of distressed companies the model misses. The AUC evaluates the model’s probability of ranking a randomly selected distressed company higher than a randomly selected non-distressed company. The main strength of the AUC measure is its robustness against imbalanced classes.

To examine the role of text and speech emotions and FinBERT-based sentiment, we started with experiments restricted to these variables only, that is, after excluding financial indicators. For this set of experiments, random forest was utilised because it is considered a state-of-the-art benchmark method in financial distress prediction [45, 54]. Existing text sentiment features were used for comparative purposes, including the sentiment score [66] and sentiment indicators from the L&M dictionary [32]. Furthermore, BERT language model [30] was used as the state-of-the-art linguistic approach. On the one hand, given the low dimensionality of our text representation, one cannot expect to beat BERT. On the other hand, our intention was to create an easily interpretable model, which is difficult in the case of the BERT model (each layer contains vectors of

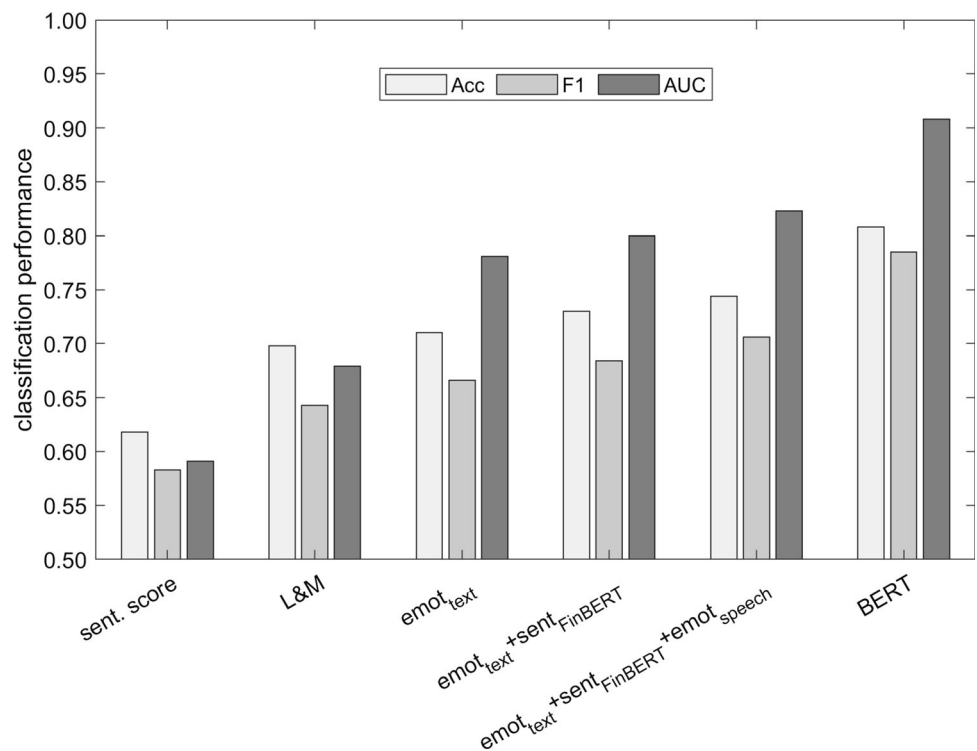
**Fig. 5** Overall financial performance of the sample in terms of Z-score. The median is denoted by the central marker, the box indicates the interquartile range, and the whiskers refer to the data instances that are not regarded as outliers



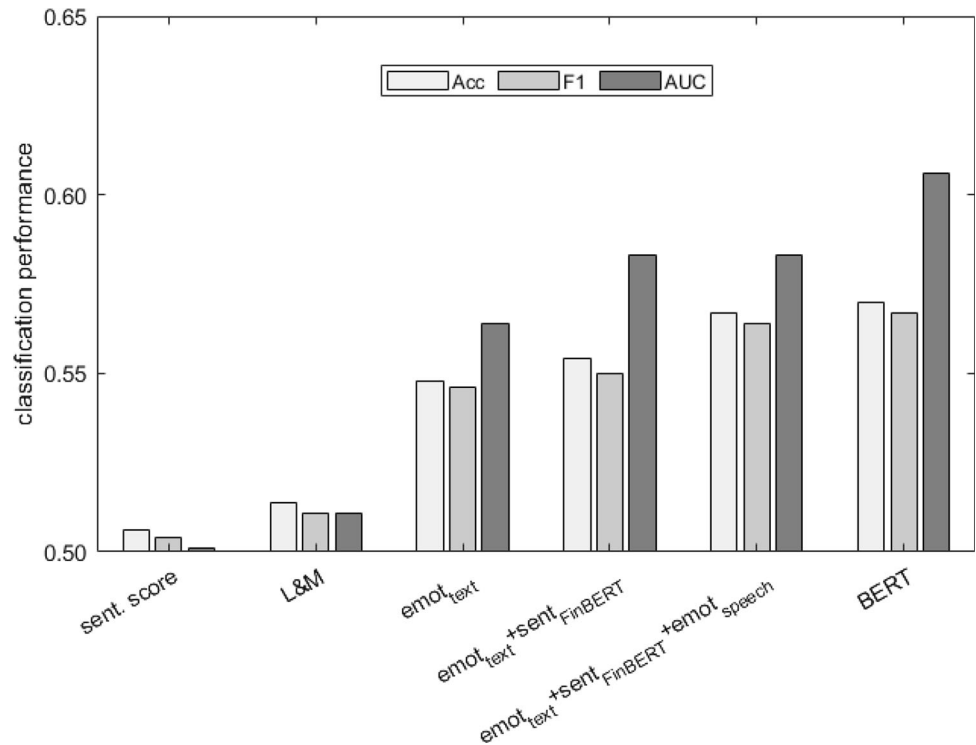
length 768). Figures 6 and 7 show the corresponding prediction performance for the class prediction and trend prediction task, respectively. First of all, the BERT-based model excelled for both classification tasks in terms of all evaluation metrics, which is consistent with existing research [30, 32]. However, our model that combined text and speech emotions with FinBERT-based sentiment was highly competitive and substantially outperformed the two

compared dictionary-based sentiment approaches. Figures 6 and 7 also reveal that the trend prediction task is too challenging for the traditional sentiment features. The results also demonstrate a considerable difference between the performance obtained by sentiment and emotional features, thus confirming the role of managerial emotions in predicting financial distress. The results in Figs. 6 and 7 also show a clear trend of increasing classification

**Fig. 6** Prediction performance for different speech and linguistic features—financial distress class prediction



**Fig. 7** Prediction performance for different speech and linguistic features—financial distress trend prediction



performance with richer representation of conference calls, validating the contribution of all proposed feature categories.

In the next step, to validate the effectiveness of the proposed prediction model, we used the proposed deep learning architecture with the above features fused with the financial indicators and compared the model's performance with a baseline approach based on financial indicators only. To further validate the prediction performance of the proposed model, a variety of previously used financial distress prediction models were used. Note that, in agreement with the original studies that introduced these models, the compared models were based solely on financial indicators 1 year in advance. The following models were considered for the comparative purposes:

- Multi-layer perceptron (MLP) [4] architecture comprising one hidden layer with 20 rectified linear units. The Adam optimiser was used to train the MLP model with the following parameters: The maximum number of epochs was 200, the learning rate was 0.001, and the L2 penalty parameter was set to 0.0001.
- XGBoost [28]. Consistent with Huang and Yen [28], the maximum tree depth was 10, the learning rate was 0.1, and the subsample ratio was 0.7.
- Stacking SVM [33]. As in Liang et al. [33], we used the SVM with linear kernels to generate the base and meta-classification models. We examined different values of

the regularisation parameter  $C = \{2^{-1}, 2^0, 2^1, \dots, 2^5\}$  for these classification models.

- Clustering-based under-sampling (CUS) + gradient boosting decision tree (GBDT) [13] with 3 clusters for CUS and 100 GBDT estimators.
- Synthetic minority over-sampling technique (SMOTE) + AdaBoost SVM (ADASVM) [58]. Consistent with [58], the total count of new minority class instances generated by SMOTE was equal to the number of majority instances minus the number of minority instances. As the SVM training parameters are not reported in the original study, different kernel functions were examined for the SVM base classifiers, and again, we experimented with the regularisation parameter  $C$  with the stacking SVM.

To implement the above comparative models, we used the Scikit-learn library and the Imbalanced-learn library.

The results in Table 4 show that the compared prediction models performed well in terms of accuracy in 1-year-ahead financial distress prediction. Among the compared models, the XGBoost model excelled, although, different from the CUS+GBDT and SMOTE+ADASVM models, it does not tackle the class imbalance problem. The outstanding performance of the baseline model indicated the advantage of larger window size (multiple time steps) in the LSTM model. We also managed to improve the prediction performance by integrating managers' sentiment and emotions. The results given in Table 4 apparently show

**Table 4** Results of 1-year-ahead financial distress prediction

Prediction model	Class prediction			Trend prediction		
	Acc	F1	AUC	Acc	F1	AUC
MLP	0.926	0.926	0.975	0.606	0.605	0.634
XGBoost	0.942	0.942	0.988	0.700	0.700	0.766
Stacking SVM	0.920	0.919	0.925	0.619	0.617	0.618
CUS+GBDT	0.934	0.934	0.982	0.661	0.660	0.710
SMOTE+ADASVM	0.920	0.920	0.967	0.603	0.603	0.653
Our baseline model with financial indicators	0.943	0.943	0.990	0.728	0.728	0.789
Our deep learning model with fused features	<b>0.954</b>	<b>0.953</b>	<b>0.992</b>	<b>0.741</b>	<b>0.741</b>	<b>0.813</b>

The best results are in bold

that financial indicators play a key role in financial distress prediction. However, the proposed sentiment and emotional features have a valuable predictive power, improving the overall prediction performance in terms of all evaluation criteria. Furthermore, the superior values of AUC suggest that the proposed prediction model performed well despite the imbalance of classes. Specifically, the accuracy of the individual classes for the imbalanced class prediction problem was as follows: 0.955 for the safe class, 0.915 for the grey zone and 0.969 for the financial distress class. These results are consistent with the findings of recent studies focusing on multi-class financial distress prediction [59], that is, it is more difficult to detect classes of financial conditions between financial health and severe financial distress. Finally, a balanced prediction performance regarding precision (0.953 for class prediction and 0.741 for trend prediction) and sensitivity (0.954 and 0.741) was obtained for both classification problems.

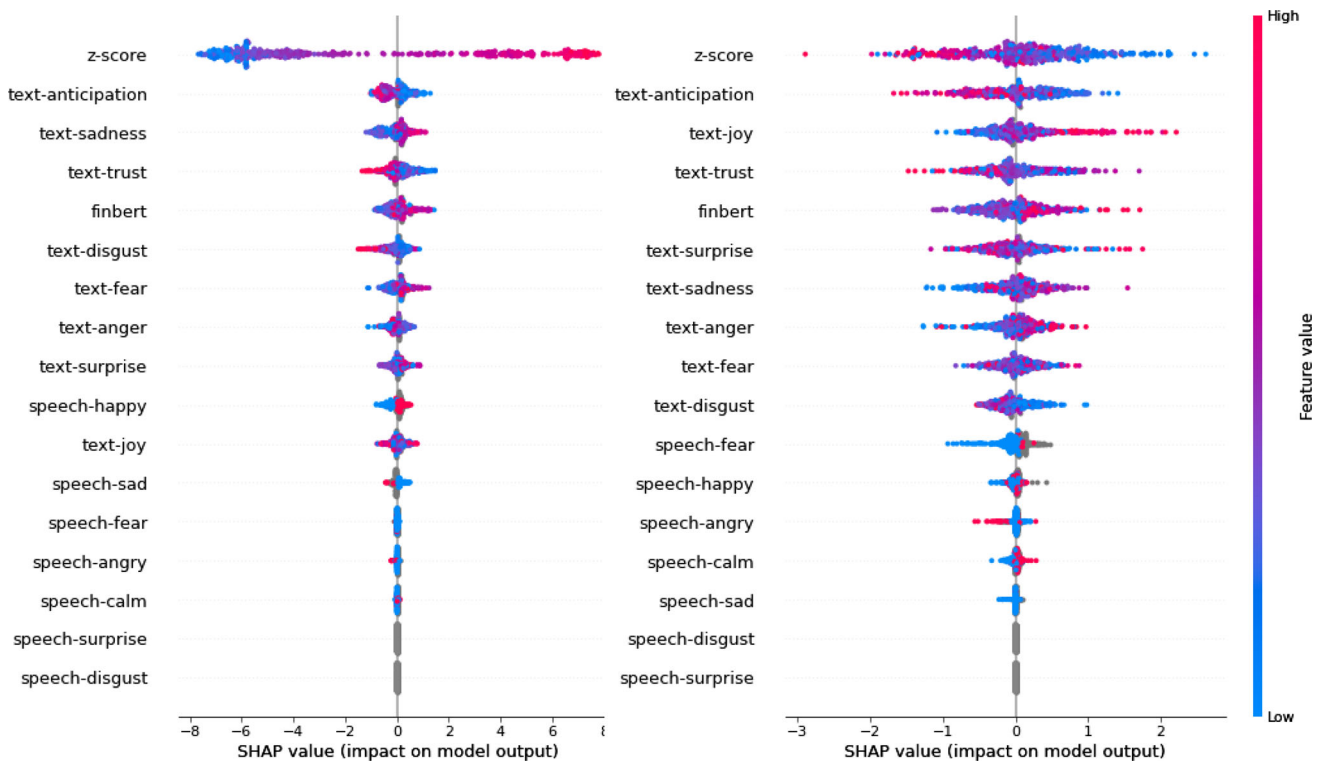
The integration of the sentiment and emotional features was even more influential in improving prediction performance in case of trend prediction. Even though the results for both classification problems were generally consistent, the problem of trend prediction appears to be a more challenging classification task. Overall, the results of the comparative models validated the established model for financial distress prediction.

SHapley Additive exPlanations (SHAP) were used to explain the effects of sentiment and emotional features on prediction results. The SHAP method is a game theoretic method that is widely used to explain the deep learning outputs [37]. A major benefit of SHAP is that it provides both local explainability (how decisions are reached for each company) and global explainability (the overall structure of the model decisions). In this study, we explored the global explainability of the prediction model to illustrate the effect of each sentiment and emotional feature on financial distress prediction. The SHAP library was used to obtain the SHAP values. Figure 8 depicts the SHAP values of the proposed features compared with the

financial indicator of greatest importance (the Z-score from the previous year). The results show that conference call transcripts were generally more relevant for the prediction model than the speech emotional states. Notably, anticipation was a crucial text emotion, suggesting that the high occurrence of forward-looking statements indicates future financial difficulties and a deterioration in the financial situation, respectively. Moreover, exciting statements (joy) were indicators of an improvement in corporate financial performance. In contrast, sadness and fear text emotions indicated financially safe companies, which can be attributed to managers' concerns about the company's financial situation and careful consideration of financial risks, respectively. Positive FinBERT-based sentiment was associated with good financial performance and its improvement in the future. The emotions of sadness and happiness were the most relevant speech emotions for financial distress prediction, while the emotions of fear and happiness were critical for the trend prediction. Speech happiness indicated a safe financial position, whereas sadness suggested financial distress. For the prediction of financial trend, the impact of happiness was similar, while the fear emotion was indicative of an increased financial performance. The angry emotion was more likely to imply deterioration. The impact of the sentiment and emotional features on model outputs is summarized in Table 5.

## 6 Conclusion

The purpose of the current study was to determine the effect of emotion and sentiment features from earnings conference calls on financial distress prediction. To this end, a dynamical deep learning-based financial distress prediction model was proposed to exploit a rich word representation and speech emotion recognition. In summary, the findings of this study suggest that both the text and speech emotions deliver valid information for predicting financial distress. The results of this study show that



**Fig. 8** SHAP values showing how sentiment and emotional features contribute to financial distress prediction (class prediction is presented on the left, and trend prediction is depicted on the right)

**Table 5** Summary of the impact of the top 5 sentiment and emotional features on model output (likelihood of financial distress or safe position)

Rank	Financial distress	Safe position
1	Anticipation (text)	Joy (text)
2	Trust (text)	FinBERT sentiment (text)
3	Surprise (text)	Sadness (text)
4	Disgust (text)	Fear (text)
5	Fear (speech)	Happy (speech)

the fusion of traditional financial indicators with the proposed sentiment and emotional indicators can improve the prediction performance. In accordance with earlier research [24, 38, 44], this study revealed that negative sentiment and emotions in audio recordings and conference call transcripts may indicate financial distress, whereas positive emotions (e.g. happiness) indicate good financial performance. The most striking result to emerge from our analysis is that managers may try to mask financial difficulties by expressing positive emotions such as anticipation and trust. In contrast, some negative emotions in managers’ statements, such as fear and sadness, show that the

company is important to managers and they care about its financial situation.

To make informed financial decisions, financial managers, investors and banks assess the financial performance of companies, which can be complex and time-consuming. Therefore, the use of accurate and interpretable models has the potential to significantly streamline this assessment. Traditional models based on financial ratios may be flawed by accounting practices or even deliberate misrepresentation of financial statements. To mitigate these problems, our model integrates financial, audio and text information, and therefore, it could be applied to reliably guide a range of financial and investment decisions.

Finally, several limitations need to be considered. The most important limitation is that only the speech emotions represented the entire audio recordings. However, the speakers’ emotional states may be subject to changes while the conference call is in progress. For instance, managerial emotions may change when discussing different areas of company’s performance. Therefore, further work needs to be done to analyse the topics of the conference call transcripts to identify topics being discussed. Another limitation is the sharp division of financial distress classes for classification purposes. Fuzzy sets could be utilised to consider the uncertainty in the overall financial performance indicators. Finally, only large companies and a

relatively short prediction horizon were considered, which provides opportunities for future research: investigating a larger sample of companies, including small- and medium-sized enterprises, and extending the prediction horizon.

**Acknowledgements** This work was supported by a grant from the Czech Sciences Foundation (No. 22-22586 S).

**Funding** Open access publishing supported by the National Technical Library in Prague.

**Data availability** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Abedin MZ, Guotai C, Hajek P et al (2022) Combining weighted smote with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex Intell Syst* 1–21
2. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun* 116:56–76
3. Alaka HA, Oyedele LO, Owolabi HA et al (2018) Systematic review of bankruptcy prediction models: towards a framework for tool selection. *Expert Syst Appl* 94:164–184
4. Alaminos D, Fernández MÁ (2019) Why do football clubs fail financially? A financial distress prediction model for European professional football industry. *PloS ONE* 14(12):e0225989
5. Altman EI, Iwanicz-Drozdowska M, Laitinen EK et al (2017) Financial distress prediction in an international context: a review and empirical analysis of Altman's z-score model. *J Int Financ Manag Account* 28(2):131–171
6. Araci D (2019) Finbert: financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*
7. Bhavan A, Chauhan P, Shah RR et al (2019) Bagged support vector machines for emotion recognition from speech. *Knowl Based Syst* 184(104):886
8. Blau BM, DeLisle JR, Price SM (2015) Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. *J Corp Finance* 31:203–219
9. Burgoon J, Mayew WJ, Giboney JS et al (2016) Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. *J Lang Soc Psychol* 35(2):123–157
10. Cao S, Jiang W, Yang B et al (2020) How to talk when a machine is listening: corporate disclosure in the age of AI. *Tech. Rep. w27950*, National Bureau of Economic Research
11. Cecchini M, Aytug H, Koehler GJ et al (2010) Making words work: using financial text as a predictor of financial events. *Decis Support Syst* 50(1):164–175
12. Charalambakis EC, Garrett I (2019) On corporate financial distress prediction: what can we learn from private firms in a developing economy? Evidence from Greece. *Rev Quant Finance Account* 52(2):467–491
13. Du X, Li W, Ruan S et al (2020) CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Appl Soft Comput* 97(106):758
14. Elend L, Tideman SA, Lopatta K et al (2020) Earnings prediction with deep learning. In: *German conference on artificial intelligence (Künstliche Intelligenz)*. Springer, pp 267–274
15. Frankel R, Jennings J, Lee J (2022) Disclosure sentiment: machine learning vs. dictionary methods. *Manag Sci* 68(7):5514–5532
16. Garain A, Ray B, Giampaolo F et al (2022) Grann: feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals. *Neural Comput Appl* 34(17):14463–14486
17. Ghazali AW, Shafie NA, Sanusi ZM (2015) Earnings management: an analysis of opportunistic behaviour, monitoring mechanism and financial distress. *Procedia Econ Finance* 28:190–201
18. Goel DP, Mahajan K, Nguyen ND et al (2023) Towards an efficient backbone for preserving features in speech emotion recognition: deep-shallow convolution with recurrent neural network. *Neural Comput Appl* 35(3):2457–2469
19. Gorgievski MJ, Bakker AB, Schaufeli WB et al (2010) Financial problems and psychological distress: investigating reciprocal effects among business owners. *J Occup Organ Psychol* 83(2):513–530
20. Hajek P (2022) Speech emotion recognition from earnings conference calls in predicting corporate financial distress. In: *ifip international conference on artificial intelligence applications and innovations*. Springer, pp 216–228
21. Hajek P, Michalak K (2013) Feature selection in corporate credit rating prediction. *Knowl Based Syst* 51:72–84
22. Hajek P, Olej V (2013) Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: *International conference on engineering applications of neural networks*. Springer, pp 1–10
23. Hajek P, Olej V (2015) Word categorization of corporate annual reports for bankruptcy prediction by machine learning methods. In: *International conference on text, speech, and dialogue*. Springer, pp 122–130
24. Hajek P, Olej V, Myskova R (2014) Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technol Econ Dev Econ* 20(4):721–738
25. Hajek P, Barushka A, Munk M (2020) Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Comput Appl* 32(23):17259–17274
26. Hobson JL, Mayew WJ, Venkatachalam M (2012) Analyzing speech to detect financial misreporting. *J Account Res* 50(2):349–392
27. Huang B, Yao X, Luo Y et al (2022) Improving financial distress prediction using textual sentiment of annual reports. *Ann Oper Res* 1–28

28. Huang YP, Yen MF (2019) A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Appl Soft Comput* 83(105):663
29. Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. *Biomed Signal Process Control* 59(101):894
30. Jiang C, Lyu X, Yuan Y et al (2022) Mining semantic features in current reports for financial distress prediction: empirical evidence from unlisted public firms in China. *Int J Forecast* 38(3):1086–1099
31. Kratzwald B, Ilić S, Kraus M et al (2018) Deep learning for affective computing: text-based emotion recognition in decision support. *Decis Support Syst* 115:24–35
32. Li S, Shi W, Wang J et al (2021) A deep learning-based approach to constructing a domain sentiment lexicon: a case study in financial distress prediction. *Inf Process Manag* 58(5):102673
33. Liang D, Tsai CF, Lu HYR et al (2020) Combining corporate governance indicators with stacking ensembles for financial distress prediction. *J Bus Res* 120:137–146
34. Livieris IE, Stavroyiannis S, Iliadis L et al (2021) Smoothing and stationarity enforcement framework for deep learning time-series forecasting. *Neural Comput Appl* 33(20):14021–14035
35. Livingstone SR, Russo FA (2018) 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
36. Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *J Finance* 66(1):35–65
37. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30:1–10
38. Mai F, Tian S, Lee C et al (2019) Deep learning models for bankruptcy prediction using textual disclosures. *Eur J Oper Res* 274(2):743–758
39. Matin R, Hansen C, Hansen C et al (2019) Predicting distresses using deep learning of text segments in annual reports. *Expert Syst Appl* 132:199–208
40. Mayew WJ, Venkatachalam M (2012) The power of voice: managerial affective states and future firm performance. *J Finance* 67(1):1–43
41. McFee B, Raffel C, Liang D et al (2015) librosa: audio and music signal analysis in python. In: *Proceedings of the 14th python in science conference*, pp 18–25
42. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. *Comput Intell* 29(3):436–465
43. Myskova R, Hajek P (2020) Mining risk-related sentiment in corporate annual reports and its effect on financial performance. *Technol Econ Dev Econ* 26(6):1422–1443
44. Nguyen BH, Huynh VN (2022) Textual analysis and corporate bankruptcy: a financial dictionary-based sentiment approach. *J Oper Res Soc* 73(1):102–121
45. Papouškova M, Hajek P (2019) Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decis Support Syst* 118:33–45
46. Price SM, Doran JS, Peterson DR et al (2012) Earnings conference calls and stock returns: the incremental informativeness of textual tone. *J Bank Finance* 36(4):992–1011
47. Price SM, Seiler MJ, Shen J (2017) Do investors infer vocal cues from CEOs during quarterly REIT conference calls? *J Real Estate Finance Econ* 54(4):515–557
48. Qi P, Zhang Y, Zhang Y et al (2020) Stanza: a python natural language processing toolkit for many human languages. In: *Association for computational linguistics system demonstrations*. ACL, pp 1–8
49. Sawhney R, Aggarwal A, Khanna P et al (2020) Risk forecasting from earnings calls acoustics and network correlations. In: *INTERSPEECH*, pp 2307–2311
50. Sawhney R, Khanna P, Aggarwal A et al (2020) Voltage: volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp 8001–8013
51. Sawhney R, Mathur P, Mangal A et al (2020) Multimodal multi-task financial risk forecasting. In: *Proceedings of the 28th ACM international conference on multimedia*, pp 456–465
52. Sawhney R, Aggarwal A, Shah R (2021) An empirical investigation of bias in the multimodal analysis of financial earnings calls. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp 3751–3757
53. Shahin I, Nassif AB, Nemmour N et al (2021) Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments. *Neural Comput Appl* 33(23):16033–16055
54. Shen F, Liu Y, Wang R et al (2020) A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. *Knowl Based Syst* 192(105):365
55. Son H, Hyun C, Phan D et al (2019) Data analytic approach for bankruptcy prediction. *Expert Syst Appl* 138(112):816
56. Sun J, Li H, Huang QH et al (2014) Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowl Based Syst* 57:41–56
57. Sun J, Fujita H, Chen P et al (2017) Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowl Based Syst* 120:4–14
58. Sun J, Li H, Fujita H et al (2020) Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf Fusion* 54:128–144
59. Sun J, Fujita H, Zheng Y et al (2021) Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. *Inf Sci* 559:153–170
60. Tang X, Li S, Tan M et al (2020) Incorporating textual and management factors into financial distress prediction: a comparative study of machine learning methods. *J Forecast* 39(5):769–787
61. Throckmorton CS, Mayew WJ, Venkatachalam M et al (2015) Financial fraud detection using vocal, linguistic and financial cues. *Decis Support Syst* 74:78–87
62. Wang G, Ma J, Chen G et al (2020) Financial distress prediction: regularized sparse-based random subspace with ER aggregation rule incorporating textual disclosures. *Appl Soft Comput* 90(106):152
63. Yang L, Ng TLJ, Smyth B et al (2020) Hml: hierarchical transformer-based multi-task learning for volatility prediction. In: *Proceedings of the web conference*, pp 441–451
64. Yang Y, Uy MCS, Huang A (2020b) Finbert: a pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*
65. Zeng Y, Mao H, Peng D et al (2019) Spectrogram based multi-task audio classification. *Multimed Tools Appl* 78:3705–3722
66. Zhao S, Xu K, Wang Z et al (2022) Financial distress prediction by combining sentiment tone features. *Econ Model* 106(105):709