INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Risk Forecasting from Earnings Calls Acoustics and Network Correlations

Ramit Sawhney[1*], *Arshiya Aggarwal*[2*], *Piyush Khanna*[3*], *Puneet Mathur*[4*], *Taru Jain*[3*], *Rajiv Ratn Shah*[3]

[1]Netaji Subhas Institute of Technology, India
[2]Adobe, India
[3]MIDAS, IIIT-Delhi, India
[4]University of Maryland, United States of America

ramits.co@nsit.net.in, arsaggar@adobe.com, piyushkhanna_bt2k17@dtu.ac.in, puneetm@cs.umd.edu, jaintaru@ieee.org, rajivratn@iiitd.ac.in

Abstract

Stock volatility is a degree of deviations from expected returns, and thus, estimates risk, which is crucial for investment decision making. Volatility forecasting is complex given the stochastic nature of market microstructure, where we use frenzied data over various modalities to make temporally dependent forecasts. Transcripts of earnings calls of companies are well studied for risk modeling as they offer unique investment insight into stock performance. Anecdotal evidence shows company CEO's vocal cues could be indicative of the stock performance. The recently developing body of work on analyzing earnings calls treat stocks as independent of each other, thus not using rich relations between stocks. To this end, we introduce the first neural model that employs cross inter-modal attention for deep verbal-vocal coherence and accounts for stock interdependence through multi-layer network embeddings. We show that our approach outperforms state-of-the-art methods by augmenting speech features with correlations from text and stock network modalities. Lastly, we analyse the components and financial implications of our method through an ablation and case study.

1. Introduction

Volatility is a statistical measure representing the dispersion of the returns of publicly traded stocks.[1] Stock volatility represents the magnitude of price swings and often models the risk associated with a stock [1]. Thus, more volatile stocks are considered riskier and present high risk-reward opportunities [2]. Earnings conference calls that are recurring events where publicly traded companies' Chief Executive Officers (CEO) prognosticate company performance presents one such high risk-reward scenario. Comprising of performance disclosure followed by a spontaneous question-answer session with financial analysts, these calls present new unique knowledge that brings significant stock price movements [3]. Despite the rich information they provide, earnings calls remain relatively underexplored, particularly from the perspective of acoustics and speech. Audio features are strongly correlated to the verbal message said by a CEO and are indicative of the speaker's emotional and affective state [4, 5]. Vocal cues and their interplay with text can help better analyze the impact earnings calls may have on financial markets.

Conventionally risk forecasting has relied on financial data [6, 7]; however, with unmatched advances in deep learning, there is a growing body of literature analyzing textual content from earnings calls and financial disclosures [8]. More recently, the use of audio processing for earnings calls has gained an interest in both financial and speech research [5]. Leveraging the audio and text data from earnings calls, MDRM [3], and HTML [9] validate the premise of speech processing for volatility prediction. However, these methods stress more on textual elements and simplistic fusion techniques for vocal cues. Additionally, they do not factor in stock interdependence.

In this paper, we build on established knowledge from financial research and recent advances in acoustics to present the first neural model that jointly exploits audio, textual, and stock correlation network information for volatility estimation. We employ inter-modal multi-utterance attention mechanisms to enhance fusion across these modalities across utterances and context. Through correlation networks created from knowledge graphs of publicly-traded companies in the S&P 500 index, our approach outperforms state-of-the-art approaches for volatility prediction by capturing inter stock relations. Through our approach, we present the below contributions:

Model: An architecture that exploits audio, text, and graph modalities by jointly learning associations through attentive mechanisms from earnings conference calls for financial risk estimation. Through comparative and ablation studies, we show the utility of augmenting vocal cues with other feature types.

Practical Applicability: Through a case study, we highlight the practical impact our method has in the financial domain and the effect of verbal-vocal interplay for analyzing earnings calls.

1.1. Related Work and Limitations of Current Approaches

There has been an abundant study for volatility prediction using historical financial data [10, 11], and numerical data beyond finance [12]. While recent work focuses on using different forms of data, there exist limitations and underexplored avenues to enhance current methods which we describe as follows:

Lack of utilizing speech features: Newer studies based on the Efficient Market Hypothesis [13] highlight the success of multimodal data in finance [14], as they capture a broader set of affecting data. Recent work uses textual data such as social media posts, news reports, etc. [15, 16], but do not analyze speech, thus not leveraging the interplay across text and audio.

Poor generalizability to high-risk macro events: The majority of existing approaches do not focus on highly volatile and macro activities such as earnings calls, where the market microstructure is highly uncertain [17]. Thus, making prediction tasks tough and risk-oriented [18].

Not all modalities play an equal role: Newer studies [3] il-

* Authors contributed equally

[1]https://www.investopedia.com/terms/v/volatility.asp

http://dx.doi.org/10.21437/Interspeech.2020-2649

lustrate the gains obtained by using vocal cues from the CEO's earnings conference calls for volatility prediction. Proposing MDRM, a late fusion model using GloVe [19] embeddings for text and hand-crafted audio features with BiLSTMs, they fuse text and speech, improving performance. Although the inclusion of both modalities enhances performance, not all modalities contribute equally. Noise in one modality can be detrimental in such multimodal frameworks. To address this, we focus on inter-modal attention across modalities and utterances.

Assuming inter-stock movement independence: HTML [9] is a transformer-based model that uses BERT [20] for textual modeling, and the same hand-crafted features as MDRM in an early fusion formulation. MDRM, HTML, and other non-speech based approaches assume no inter-dependence between stocks and do not exploit rich correlations across stocks. Methods employing graphs to use inter-stock relations also show the potency of exploiting such relations [21].

2. Context and Problem Formulation

Measuring stock volatility: Following [9, 3], for a given stock, with a close price of p_i on trading day i, we use Equation 1 to calculate the average volatility over n days following the earnings call.

$$v_{[0,n]} = \ln\left(\sqrt{\frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n}}\right) \tag{1}$$

where, the return price r_i is defined as $\frac{p_i}{p_{i-1}} - 1$.

Formulation: Given an earnings call c, comprising of an audio component A, and the corresponding aligned text component T, we aim to learn a function $f(c_{\{T,A\}}) \rightarrow v_{[0,n]}$. Following [3, 9], we experiment with $n \in \{3, 7, 15, 30\}$ days to analyze the performance over both short and long term periods.

3. Method

3.1. Modeling Vocal Cues: Audio Feature Extraction

Motivation: Driven by extensive studies [4, 5] on the correlation of the psychological state of a speaker with different acoustic features, we extend the feature sets of previous works [3, 9]. These features include 11 point Amplitude Perturbation Quotient (APQ 11) Shimmer and DDA Shimmer, which are linked to stress and anxiety [22, 23]. The ratio of voiced to unvoiced frames in audio is obtained, which is indicative of the pace at which a person speaks and reflects inconsistencies between verbal and vocal cues [24, 25]. We extracted a total of 26 features from each audio utterance using Praat [26].

Formulation: Following [3] we employ the Iterative Forced Alignment (IFA) algorithm to segment and align each utterance of the transcript with the audio utterance. We represent the segmented audio clips as $(a_1, a_2, ..., a_n)$ where $a_i \in \mathbb{R}^n$, n being the number of clips of an earnings call, with each clip being represented by 26 acoustic features. We utilize a BiLSTM layer that encodes these features as shown by Equation 4.

$$\overrightarrow{A_t^{(f)}} = BiLSTM^{(f)}(a_t, A_{t-1}^{(f)}) \tag{2}$$

$$\overleftarrow{A_t^{(b)}} = BiLSTM^{(b)}(a_t, A_{t+1}^{(b)}) \tag{3}$$

$$A_t = [\overrightarrow{A_t^{(f)}}, \overleftarrow{A_{T-t}^{(b)}}] \tag{4}$$

3.2. Modeling Verbal Cues: Sentence Encoding

Motivation: To leverage contextual attributes of the earnings calls, we extract textual features from sentences in the transcripts. We use Siamese BERT networks [27] as a sentence encoder, that builds on BERT [20] to perform semantic similarity assessment. Siamese BERT[2] fine-tunes sentence embeddings that help to capture context across the transcript better.

Formulation: We represent the sentences of each clip in an earnings call as $(t_1, t_2, ..., t_n)$ where $t_i \in \mathbb{R}^n$, n being the number of sentences. We encode these as:

$$s_i = SiameseBERT(t_i) \tag{5}$$

In order to fine-tune BERT, the Siamese BERT networks optimize on the triplet loss objective, so as to produce sentence embeddings that are semantically meaningful. These resultant intermediate representations s_i are encoded through a BiLSTM layer, and we obtain a text encoding T_t, similar to Equation 4.

3.3. Cross-Modal Gated Attention Fusion

Vocal cues play a dual role in examining the validity of speech and understanding the context of spoken sentences. To leverage the interplay of verbal and vocal cues, we apply a Cross-Modal Gated Attention Fusion mechanism that attends over the contextual utterances. The mechanism computes correlations among text and audio modalities of the target utterance and its contextual neighbors. Such associations help to identify and select the most relevant modality over each contextual utterance window. Inspired by [28], we utilize the gated attention mechanism shown in Figure 1 to generate modality-specific attentive representations.

Formulation: The attention mechanism [29] captures cross-modal information from audio and text encoding by computing the correlation matrices for the audio and text modalities $C_t, C_a \in \mathbb{R}^{n \times n}$ as shown by Equation 6. Further, the attention weights over the correlation matrices are computed using the softmax activation to get contextual inter-modal matrices W_t, W_a that capture the contextual dependencies in the utterances (Equation 7). Subsequently, we compute the modalitiy-wise attentive representations G_t, G_a. Finally, a multiplicative gating mechanism is introduced to attend the important components of text and audio sequences to get the final attentive feature embeddings F_t, F_a which are concatenated as:

$$C_t = AT^T,\ C_a = TA^T \tag{6}$$

$$W_a = softmax(C_a),\ W_t = softmax(C_t) \tag{7}$$

$$G_a = W_a \cdot T,\ G_t = W_t \cdot A \tag{8}$$

$$F_a = G_a \odot A,\ F_t = G_t \odot T \tag{9}$$

where $\cdot$ represents the dot product and $\odot$ represents the element wise multiplication.

3.4. Augmenting Speech with Network correlations

Existing methods typically treat stocks as independent of each other and ignore their interdependence. However, the rich inter-dependencies between stocks (companies) contain valuable clues for financial modeling tasks [21, 30]. Learning representations over stock relations can improve volatility forecasting. Following [21], we perform graph-based learning by using two relation networks shown in Table 1, described as follows:

[2]Implementation used: `https://github.com/UKPLab/sentence-transformers`

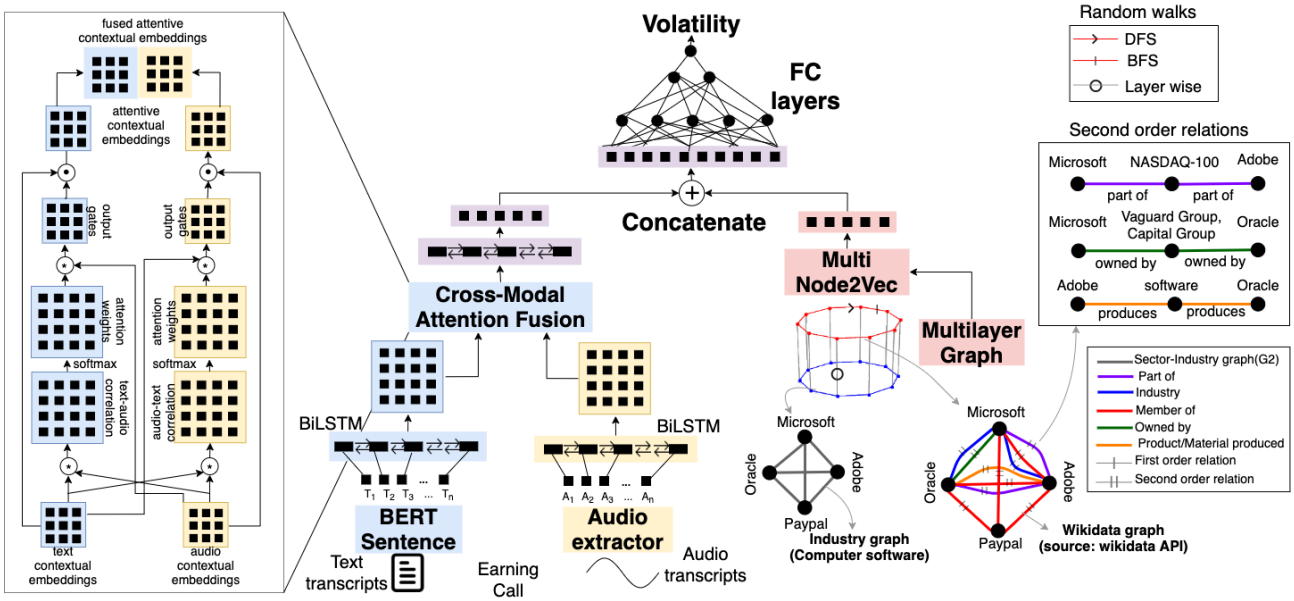


Figure 1: *Schematic diagram for model architecture.*

Table 1: *Stock-Sector and Wiki-Company graph statistics*

Graph	$\|V\|$	$\|E\|$	Degree_{avg}
Stock-Sector	277	426	3.07
Wiki-Company	277	5,476	39.54

Stock-Sector Relations: Stocks belonging to the same sector are similarly influenced by the prospects of that industry. Hence, we extract the company taxonomy structure of NASDAQ and NYSE. The sector-industry graph $G_{SS} = (V, E_{SS})$ where V is the set of all stocks in S&P 500 in 2017 and $e \in E_{SS}$ if two stocks $v_1, v_2 \in V_{SS}$ are part of the same industry.

Wiki-Company based Relations: Connections between companies and their relational entities are mined from WikiData[3] knowledge base. We extract first and second-order relations, mentioned in the Appendix of [21]. The Wiki-Company graph $G_{WC} = (V, E_{WC})$ is built on the same set of vertices as that of Stock-Sector graph with edges present between two companies (objects) if the same entity (subject) acts as a relation bridge (predicate) between them.

Multi-node2vec: We represent the two stock relation networks as a multi-layer graph. We use Multi node2vec algorithm[4] [31] that fuses multiple unordered graphs as multilayer networks to learn representations for these networks. The network G_N^2 thus formed is homogeneous in vertices ($N = 277$ for both graphs) but heterogeneous in edges. The graph realization problem boils down to a joint likelihood maximization algorithm as given by Equation 10 for any node u where $v = Ne(u)$ is its node neighbours and f_w is the feature representation of any node w.

$$\mathcal{L} = \sum_{u \in \mathcal{N}} \sum_{v \in Ne(u)} [f_v^T f_u - \log(\sum_{w \in \mathcal{N}} \exp(f_w^T f_u)] \qquad (10)$$

For each unique node $n \in N$, the algorithm performs a neighbourhood search across both layers, for which it uses an inter-layer walk parameter r, in addition to the regular return parameter p, and in-out parameter q used in [32]. For every pair of corresponding nodes in the two layers, a default edge is created to account for cross-layer relations.

3.5. Multimodal Fusion

The output from the cross-modal attention fusion of audio encoding A_t and text encoding T_t is passed through another BiLSTM, and its contextual output H_t is concatenated with graph embeddings (G_N^2), which is then passed through a fully connected layer ϕ. The resultant output $\hat{z_{reg}}$ of the proposed model is used for regressing volatility values as illustrated by Equation 11, and optimized over Mean Squared Error (MSE).

$$\hat{z_{reg}} = \phi(W^T[concat(H_t, G_N^2)]) \qquad (11)$$

4. Dataset

We used the S&P 500 2017 Earnings Conference Calls dataset [3] for all experiments. The dataset consists of 562 earnings call audio recordings and their transcripts for 274 companies in the S&P 500 index[5]. Each call is segmented into a sequence of audio clips aligned with their corresponding text sentences, as spoken by the CEO during the call, summing up 88,829 aligned sentences. We temporally divide the data into train, validation and, test sets in a ratio of 70 : 10 : 20, respectively, in chronological order to ensure future data, is not used for forecasting. We extract stock prices for each company using Yahoo Finance[6] from 1 January 2017 till 31 December 2017.

[3] https://www.mediawiki.org/wiki/Wikibase/DataModel/JSON

[4] Implementation used: https://github.com/jdwilson4/multi-node2vec

[5] We were unable to map price data for 11 data points, which were subsequently dropped.

[6] https://finance.yahoo.com/

Table 2: *n-day volatility prediction errors for models*

Model	MSE_{avg}	MSE_3	MSE_7	MSE_{15}	MSE_{30}
V_{past}	1.12	2.99	0.83	0.42	0.23
LSTM [34]	0.75	1.97	0.46	0.32	0.24
HAN(GloVe) [35]	0.60	1.43	0.46	0.31	0.20
MDRM (Audio) [3]	0.60	1.41	0.44	0.32	0.22
MDRM [3]	0.58	1.37	0.42	0.30	0.22
HTML (Text) [9]	0.46	1.18	0.37	**0.15**	0.13
HTML [9]	0.40	0.85	0.35	0.25	0.16
Ours	**0.35**	**0.73**	**0.33**	0.22	**0.12**

5. Experimental Settings

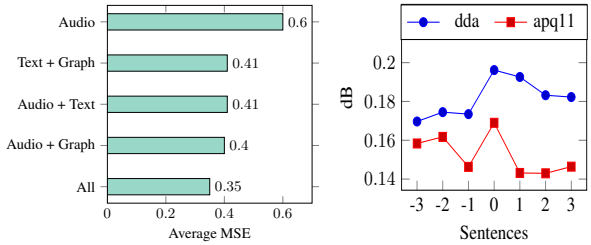
Training Setup: We explored the following hyperparameters: number of hidden layers, size of hidden layers of BiLSTM and Dense, dropout $\delta \in [0, 0.8]$, learning rate $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, batch size $b \in \{8, 16, 32, 64\}$ and epochs (< 100). Adam [33] was employed for optimizing the MSE of the proposed model. All three BiLSTM's used after text, audio, and cross-modal gated attention fusion were set to have 100 hidden units each. The number of neurons in the time distributed dense layer following the audio and text BiLSTM's is 100, while the number of neurons in the penultimate dense layer is 50. SiameseBERT outputs a 768-dimensional embedding for each sentence where 768 is the size of the hidden layer dimension in the BERT architecture. The maximum number of audio clips in any call is 520. Hence, all audio and textual input features lesser than maximum length are padded. The node neighborhood search procedure of the multi-node2vec algorithm depends on three hyper-parameters- p, q, and r - that dictate the exploration of the random walk away from the source node and the tendency to traverse layers. We use the default values of $p = 1$, $q = 0.50$, and $r = 0.25$ to traverse breadth-wise, depth-wise, and across the layers with appropriate probabilities, thus capturing all possible correlations.

Baselines and Evaluation metrics: We compare our approach with previously studied conventional methods and recent state-of-the-art approaches against the MSE between the true and predicted log volatility. Following [3], we use V_{past}: a measure of past volatility, which could be indicative of future volatility. We also compare against LSTM [34]-based approaches that use historical price data, and Hierarchical Attention Networks [35] that are commonly used for analyzing earnings calls transcripts. We compare against the previously discussed deep multimodal architectures: MDRM [3] and HTML [9], which are the current state-of-the-art. We also consider their unimodal variants MDRM (Audio) and HTML (Text).

6. Results & Analysis

6.1. Drawing Insights from Quantitative Comparisons

In Table 2, we report quantitative comparisons against the baselines discussed above. Sudden drift in volatility following an earnings call makes short term volatility estimation more important and chaotic [36]. It is evident from the results that our approach gives a significant gain for these complex tasks ($\tau = 3, 7$ days). Based on Post Earnings Announcement Drift (PEAD) [36], similar to works [9, 3], we observe diminishing gains in comparison to simple baselines as time elapses from the calls. Through the ablation study shown in Figure 2a, we find the performance gains by augmentation of verbal cues and correlation networks with speech. The best results achieved by fusing all three modalities can be attributed to the fusion of speech and text via attention mechanism, and network embeddings. Thus



(a) *Ablation Analysis* (b) *DG: Shimmer Analysis*

Figure 2: *Qualitative insights: Ablation and Case study*

diverse contributing factors across modalities boost prediction.

6.2. Case Study: Impact of Vocal Cues and Stock Networks

We conduct a case study to analyze the significance of the multimodal components of our approach. Our analysis is based on the Q3-2017 earnings call for DG (Dollar General), an American retail chain. The stock's price became highly volatile for a few days following the earnings call. We study the audio features of the call through the CEO's vocal cues and the text transcript and correlation graphs of the company. Figure 2b shows the disparity between CEO's vocal and verbal cues around the utterance *"It is also important to note that we're lapping significant working capital improvements from 2016. We continue to be pleased with our solid cash flow generation."* Here, the colors represent token-level attention. While the language seems positive, we see a sudden spike in the shimmer features in the CEO's voice while speaking this sentence, showing disagreement across verbal and vocal cues. As per acoustic research [22], an elevated shimmer pattern can be an indicator of underlying stress in human speech. After the call, it was noted that the company's gross margin slipped by 0.4% due to the increased transportation costs due to hurricane Irma in 2017.

On analyzing relation graphs, we observe that DG has edge connections with WMT (Walmart) and TGT (Target Corp.), both of which are retail variety stores, like DG. Analysts had estimated a negative impact of about $2.8 Billion on the retail sector due to the hurricane Irma. This examination is also reflected in the high volatilities recorded for WMT and TGT during the same quarter. A unimodal model may miss these subtle disparities between text and audio.

7. Conclusion

Volatility, measured as a deviation in returns, is a reliable indicator of market risk linked with a stock. Owing to its significance across finance, and beyond, volatility forecasting has seen applications through neural architectures. A rich source of company information is earnings calls, which provide high risk-reward opportunities given their uniqueness. Although evidence shows that enriching models with speech and inter-stock correlations can improve volatility forecasting, this area is underexplored. To address this, we propose the first neural architecture that jointly exploits coherence over speech, text, and inter-stock correlations. Through experiments on S&P 500 index data, we show the merit of cross-modal gated attention fusion and graph-based learning. We analyze an earnings call of Dollar General, a US retail chain, for qualitative insight.

8. References

[1] S.-H. Poon and C. W. Granger, "Forecasting volatility in financial markets: A review," *Journal of economic literature*, vol. 41, no. 2, pp. 478–539, 2003.

[2] E. F. Fama and J. D. MacBeth, "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, vol. 81, no. 3, pp. 607–636, 1973.

[3] Y. Qin and Y. Yang, "What you say and how you say it matters: Predicting financial risk using verbal and vocal cues," in *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, 2019, p. 390.

[4] K. Fish, K. Rothermich, and M. D. Pell, "The sound of (in) sincerity," *Journal of Pragmatics*, vol. 121, pp. 147–161, 2017.

[5] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[6] W. Kristjanpoller, A. Fadic, and M. C. Minutolo, "Volatility forecast using hybrid neural network models," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2437–2442, 2014.

[7] J. Zheng, A. Xia, L. Shao, T. Wan, and Z. Qin, "Stock volatility prediction based on self-attention networks with social information," in *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE, 2019, pp. 1–7.

[8] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *arXiv preprint arXiv:2003.01859*, 2020.

[9] L. Yang, T. L. J. Ng, B. Smyth, and R. Dong, "Html: Hierarchical transformer-based multi-task learning for volatility prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 441–451.

[10] C. K. Jones, "Modern portfolio theory, digital portfolio theory and intertemporal portfolio choice," *American Journal of Industrial and Business Management*, vol. 7, pp. 833–854, 2017.

[11] I. D. Dichev and V. W. Tang, "Earnings volatility and earnings predictability," *Journal of accounting and Economics*, vol. 47, no. 1-2, pp. 160–181, 2009.

[12] J. Y. Campbell, J. J. Champbell, J. W. Campbell, A. W. Lo, A. W. Lo, and A. C. MacKinlay, *The econometrics of financial markets*. princeton University press, 1997.

[13] B. G. Malkiel, "The efficient market hypothesis and its critics," *Journal of economic perspectives*, vol. 17, no. 1, pp. 59–82, 2003.

[14] Q. Li, J. Tan, J. Wang, and H. Chen, "A multimodal event-driven lstm model for stock prediction using online news," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[15] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock price prediction using news sentiment analysis," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2019, pp. 205–208.

[16] J. Tan, J. Wang, D. Rinprasertmeechai, R. Xing, and Q. Li, "A tensor-based elstm model to predict stock price using financial news," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[17] J. L. Rogers, D. J. Skinner, and A. Van Buskirk, "Earnings guidance and market uncertainty," *Journal of Accounting and Economics*, vol. 48, no. 1, pp. 90–109, 2009.

[18] L. B. Andersen, "Efficient simulation of the heston stochastic volatility model," *Available at SSRN 946405*, 2007.

[19] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[21] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, "Temporal relational ranking for stock prediction," *ACM Transactions on Information Systems*, vol. 37, no. 2, p. 1–30, Mar 2019. [Online]. Available: http://dx.doi.org/10.1145/3309547

[22] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–1081–IV–1084.

[23] P. Mongia and R. Sharma, "Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual," *Journal of Computer Networks and Communications*, vol. 2014, 11 2014.

[24] J. Přibil and A. Přibilová, "Spectral flatness analysis for emotional speech synthesis and transformation," *Lecture Notes in Computer Science*, vol. 5641, pp. 106–115, 01 2009.

[25] M. Viswanathan, Z.-X. Zhang, X.-W. Tian, and J. S. Lim, "Emotional-speech recognition using the neuro-fuzzy network," in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC '12. New York, NY, USA: Association for Computing Machinery, 2012.

[26] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot Int*, vol. 5, pp. 341–347, 01 2001.

[27] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3973–3983.

[28] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," *arXiv preprint arXiv:1905.05812*, 2019.

[29] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," *arXiv preprint arXiv:1606.01549*, 2016.

[30] X. Guo, H. Zhang, and T. Tian, "Development of stock correlation networks using mutual information and financial big data," *PloS one*, vol. 13, p. e0195941, 04 2018.

[31] J. D. Wilson, M. Baybay, R. Sankar, and P. E. Stillman, "Fast embedding of multilayer networks: An algorithm and application to group fmri," *CoRR*, vol. abs/1809.06437, 2018.

[32] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 855–864.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[34] X. Jiang, "Bitcoin price prediction based on deep learning methods," *Journal of Mathematical Finance*, vol. 10, pp. 132–139, 01 2020.

[35] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489. [Online]. Available: https://www.aclweb.org/anthology/N16-1174

[36] V. L. Bernard and J. K. Thomas, "Post-earnings-announcement drift: Delayed price response or risk premium?" *Journal of Accounting Research*, vol. 27, pp. 1–36, 1989. [Online]. Available: http://www.jstor.org/stable/2491062