

AMA-LSTM: Pioneering Robust and Fair Financial Audio Analysis for Stock Volatility Prediction

Shengkun Wang¹, Taoran Ji², Jianfeng He¹, Mariam Almutairi¹,
Dan Wang³, Linhan Wang¹, Min Zhang¹, Chang-Tien Lu¹

¹Department of Computer Science, Virginia Tech

²Department of Computer Science, Texas A&M University - Corpus Christi

³School of Business, Stevens Institute of Technology

{shengkun, jianfenghe, malmutairi, linhan, minzhang, ctlu}@vt.edu

taoran.ji@tamucc.edu, dwang35@stevens.edu

Abstract

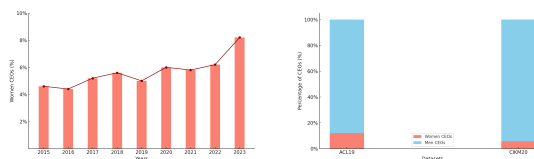
Stock volatility prediction is an important task in the financial industry. Recent advancements in multimodal methodologies, which integrate both textual and auditory data, have demonstrated significant improvements in this domain, such as earnings calls¹. However, these multimodal methods have faced two drawbacks. First, they often fail to yield reliable models and overfit the data due to their absorption of stochastic information from the stock market. Moreover, using multimodal models to predict stock volatility suffers from gender bias and lacks an efficient way to eliminate such bias. To address these aforementioned problems, we use adversarial training to generate perturbations that simulate the inherent stochasticity and bias, by creating areas resistant to random information around the input space to improve model robustness and fairness. Our comprehensive experiments on two real-world financial audio datasets reveal that this method exceeds the performance of current state-of-the-art solution. This confirms the value of adversarial training in reducing stochasticity and bias for stock volatility prediction tasks².

1 Introduction

In the stock market, predicting the exact price of a stock is deemed impossible (Nguyen et al., 2015), yet it is widely accepted within the financial industry that one can forecast a stock’s volatility level using publicly available information (Dumas et al., 2009). Stock price volatility, defined as the standard deviation of a stock’s returns over a specific period, serves as a commonly used indicator of financial risk. In the past, research efforts have focused on employing time-series models, using

¹Earnings calls are public available and often involve the management team of a public company and interested parties to discuss the company’s earnings.

²Code are available at: <https://github.com/haol1zhao/AMA-LSTM>



(a) S&P 500 Women CEO (b) CEO Gender of Datasets

Figure 1: (a) displays the percentage of woman CEO in recent years, and (b) compares the proportion of female to male CEO within the two datasets utilized.

historical stock prices, to predict future volatility (Kristjanpoller et al., 2014). In recent years, advancements in Natural Language Processing (NLP) and Speaker Recognition (SR) have opened up new possibilities for the task. For example, researchers leverage novel sources of textual and audio data, including financial news (Yang et al., 2018), financial reports (Rekabsaz et al., 2017), social media (Wang et al., 2023), earnings call (Qin and Yang, 2019) and merger and acquisition (M&A) call (Sawhney et al., 2021b) to predict stock volatility. These diverse data sources provide a richer, more nuanced understanding of market dynamics and investor sentiment, significantly enhancing the prediction of stock volatility.

However, the inherent nature of financial data presents two unique challenges. First, textual data is discrete and stochastic (Yuan et al., 2021), and audio signals possess a high temporal resolution (Donahue et al., 2018). This characteristic makes both text and audio vulnerable to adversarial examples, which can effortlessly mislead human evaluators (Carlini and Wagner, 2018; Xie et al., 2022). Second, the financial data is biased. Research (Das et al., 2021) has delved into various biases within this field, with gender imbalance being a particularly prominent issue. Bigelow et al. (2014) found that female CEOs, despite having comparable credibility, are perceived as less capable of attracting capital. In the context of earnings calls, Comprix

et al. (2022) observed a significant underrepresentation of female executives. This biased financial audio data is further exacerbated by deep learning models, which tend to amplify variations in audio features due to the scarcity of female training examples, leading to gender bias (Sawhney et al., 2021a).

To address the aforementioned issues, we use adversarial training to enhance the robustness and fairness of financial data interpretation models. Adversarial training was initially used in computer vision tasks (Goodfellow et al., 2014), and then expanded to NLP (Miyato et al., 2016) and SR (Sun et al., 2018) fields. Our method differs from the feature space adversarial training (Feng et al., 2019). It enhances model robustness and reduces output bias by introducing adversarial examples directly from input embeddings, making the approach better aligned with the privacy and proprietary constraints of financial modeling.

We have applied our adversarial training method to a multimodal attentive LSTM model, which efficiently processes information from two modalities. By introducing perturbations into the input space and dynamically optimizing these perturbations to maximize their impact on the model’s output, we enhance the model’s ability to perform well on both clean and perturbed data. The adversarial training method effectively guarantees stable training and robust performance of the model when dealing with stochastic and biased financial data. This training approach can also apply to the Transformer, but we’ve found the attentive LSTM to be more effective for our scenario. This is because the Transformer struggles with temporal information and lacks enough data to learn its many parameters effectively. To the best of our knowledge, this work is the first one to explore the potential of adversarial training in multimodal learning financial audio task. The main contributions of this paper are summarized as:

- We suggest a method of adversarial training that tackles the issues of randomness and bias, and implement it on a deep learning model for stock volatility prediction.
- We study the impact of gender bias in earnings call audio on stock volatility predictions by training on gender-specific audio features.
- We delve into the stochastic nature of financial data, emphasizing the critical need to address the randomness inherent in input features.
- We test our method on two public benchmarks,

showing it outperforms strong baselines and proving that adversarial learning increases its robustness and fairness.

2 Related Work

Our work is closely related with the following three lines of research:

Stock Volatility with Multimedia Data : Traditional stock volatility prediction method relies on historical pricing data and typically employs both continuous and discrete time-series models (Idrees et al., 2019). In addition to stock prices, financial news, analyst reports, earnings reports, and social media have been proven to significantly enhance stock prediction tasks (Wang et al., 2023; Zhang et al., 2021; Wang et al., 2020; Zhang et al., 2018; Liu et al., 2017; Rekabsaz et al., 2017). Furthermore, recent multimodal models that employ LSTM (Qin and Yang, 2019), GCN (Sawhney et al., 2020) and Transformer (Yang et al., 2020) to extract features from earnings calls audio and combine it with textual transcript, have achieved state-of-the-art results in stock volatility prediction.

Bias in Financial Audio Data : Financial audio features can indicate a speaker’s emotional and psychological state (Fish et al., 2017). Previous research has demonstrated that audio features, such as pitch and intensity, differ significantly across genders (Burris et al., 2014). Especially in earnings calls, female executives are highly underrepresented, Suresh and Guttag (2019) noted that only 5% of Fortune-500 CEOs are women. Moreover, under identical conditions, men are often perceived as more charismatic than female executives (Novák-Tót et al., 2017). This disparity in audio features becomes more pronounced in deep learning models due to the scarcity of female training examples, leading to the manifestation of gender bias.

Adversarial Training : The concept of adversarial training is quite direct, it enhances training data by incorporating adversarial examples in each iteration of the training process. Initially brought to the forefront by Szegedy et al. (2013), this idea involved training neural networks with a combination of adversarial and clean examples (Goodfellow et al., 2014). Huang et al. (2015) further developed this concept by framing adversarial training as a min-max optimization problem. Adversarial training is now widely recognized as one of the most effective methods for boosting the robustness of deep learning models (Athalye and Carlini, 2018).

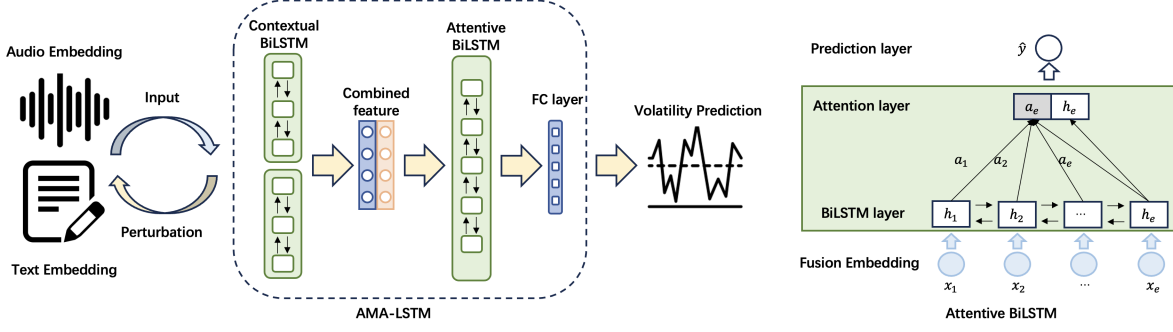


Figure 2: Illustration of the adversarial multimodal attentive LSTM architecture and an attentive BiLSTM block.

The existing works mainly concentrate on computer vision tasks (Goodfellow et al., 2014). Due to the simplicity and effectiveness, adversarial training have recently extended its application to NLP (Zang et al., 2019) and SR (Mei et al., 2022). In these areas, researchers generate adversarial examples from input embeddings, such as word and audio embeddings, to create perturbation-resistant areas around the input space.

3 Problem Formulation

The stock volatility prediction problem is formulated following (Kogan et al., 2009). For a given stock with an adjusted closing price p_i on trading day i , we calculate the average log volatility over n days following the earnings call as follows:

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

where r_i is the stock return on day i and, \bar{r} represents the average stock return across a n days window. We define the return as $r_i = \frac{p_i}{p_{i-1}} - 1$. Based on the value of n , the horizon of the prediction can be adjusted. Here, we consider various window sizes n , such as 3, 7, and 15 trading days. In our study, we have P earnings call transcripts and the longest one has Q sentences. We denote $x = \{x_1, \dots, x_e\}$ as the multimodal fusion embedding and x_e is the fusion embedding for e -th earnings call where $x_e \in \mathbb{R}^{Q \times D}$. This embedding x_e encompasses an audio component $A_e \in \mathbb{R}^{Q \times D_a}$, corresponding aligned text component $T_e \in \mathbb{R}^{Q \times D_t}$, where D_a and D_t denote the dimensions of audio and text embeddings, respectively. The total dimension D is the concatenation of these two. Our goal is to develop a predictive regression function $f(x_e) \rightarrow v_{[0,n]}$.

4 Adversarial Multimodal Attentive LSTM Architecture

Our adversarial multimodal attentive LSTM (AMA-LSTM) has two parts. We first introduce the architecture of the attentive multimodal model, which operates without adversarial training. Then, we specify the attentive LSTM block, which contains four components: feature mapping layer, LSTM layer, attention layer, and prediction layer. Secondly, we thoroughly explain how adversarial training helps multimodal models improve robustness and fairness.

4.1 Multimodal Attentive LSTM

The attentive multimodal model comprises two primary components. As shown in Figure 2, the first two contextual LSTM blocks are designed to extract unimodal features from either text or audio data. These blocks adeptly capture relationships and dependencies within each individual modality. In the second component, the extracted features from both text and audio modalities are combined and fed into an attentive LSTM block, followed by a fully-connected layer.

LSTM layer. LSTM’s (Hochreiter and Schmidhuber, 1997) proficiency in capturing long-term dependencies has made it a popular choice for processing multimodal sequential data (Qin and Yang, 2019). The core principle of LSTM involves recurrently transforming an input sequence into a series of hidden representations. At each time-step, LSTM updates the hidden representation h_e by integrating the current input x_e with the preceding hidden representation h_{e-1} , thereby capturing the sequential dependencies: $h_e = LSTM(x_e, h_{e-1})$. Adapting this concept, we employ a BiLSTM layer to better capture the bidirectional temporal patterns and sequential dependencies in text and audio fea-

tures. The layer maps the sequence $[x_1, \dots, x_e]$ into hidden representations $[h_1, \dots, h_e] \in \mathbb{R}^{U \times D}$ with the dimension of U .

Attention Layer. The attention mechanism compresses hidden representations from different time-steps into a unified overall representation and assigns adaptive weights to each step. The key concept here is the recognition that data from different time-steps may vary in their contribution to the overall sequence representation. For financial data representation, status at different time-steps might also contribute differently. As such, we use an attention mechanism to aggregate the hidden representations as:

$$a_e = \sum_{e=1}^E \alpha_e h_e, \alpha_e = \frac{\exp(c_e)}{\sum_{e=1}^E \exp(c_e)}, \quad (2)$$

$$\alpha_e = u \tanh(\mathbf{W}_a \mathbf{h}_e + \mathbf{b}_a),$$

where $W_a \in \mathbb{R}^{K \times U}$, b_a and $u \in \mathbb{R}^K$ are parameters to be learned.

Prediction Layer. Instead of predicting directly from a_e , we first involves concatenating a_e with the last hidden state h_e to form the final latent representation of the earnings call:

$$l_e = [a_e, h_e], \quad (3)$$

where $l_e \in \mathbb{R}^{2U}$. The rationale behind this is to give additional emphasis to the most recent time-step, which is often considered highly indicative of future volatility. Utilizing l_e , we then apply a fully connected layer as our predictive function. This layer is responsible for estimating the classification confidence y_e , formulated as $y_e = w_m l_e + b_m$.

4.2 Adversarial Training

Applying multimodal attentive LSTM models to forecast stock volatility presents inherent challenges due to the stochastic and biased nature of financial data, notably from earnings calls (Sawhney et al., 2021a). Earnings calls are rich with qualitative information that is often speculative and sentiment-driven, contributing to the stochastic and biased nature of the data features (Blau et al., 2015). An adversarial training approach counters this by perturbing input data to simulate these uncertainties, thereby pushing the model to maintain robustness predictions and reduce bias. This method, a strategic deviation from training solely on clean data, aims for robustness by incorporating the worst-case scenarios within its optimization function. The goal is to craft models that

are both sensitive to the nuanced dynamics of the market and resistant to overfitting, ensuring reliable performance and robustness to the financial task.

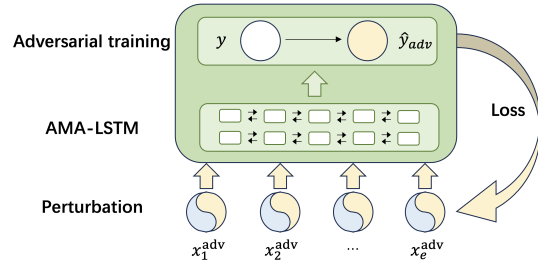


Figure 3: Illustration of the AMA-LSTM adversarial training process. Perturbations (δ) are derived by computing the gradients of the token embeddings in relation to the loss function.

As shown in Figure 3, in developing an adversarial training model for financial data, inspired by Shaham et al. (2018) we approach the problem through robust optimization. The adversarial training process seeks to solve the objective function of AMA-LSTM:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} \left[\max_{\delta \in S} (f_{\theta}(x + \delta) - y)^2 \right] + \lambda R(\theta), \quad (4)$$

where $(x, y) \sim \mathcal{D}$ denotes the training data drawn from the distribution \mathcal{D} , S defines the allowed range of perturbations and $R(\theta)$ represents the L2 regularization term, which is the sum of the squared norms of the parameters, expressed as $R(\theta) = \sum_i \|\theta_i\|^2$. This method frames adversarial training as a min-max problem, focusing on minimizing the regression error while simultaneously considering an adversary's efforts to maximize this error through input perturbations. The essence of this approach is the generation of strong adversarial examples that push the model to find the best possible parameters under worst-case scenarios. Therefore, adversarial training, by creating anti-disturbance areas, could enable the multimodal model to capture the stochastic and biased properties of earnings calls input. Since it is intractable to directly calculate δ we employed a multi-step gradient based attack method (Madry et al., 2017) for solving the problem as follows:

$$x_{adv}^{t+1} = \text{Proj}_{x+S} \left(x_{adv}^t + \beta \text{sign}(\nabla_x (f_{\theta}(x_{adv}^t) - y)^2) \right), \quad (5)$$

where t is the current step and β is the step size, it would lead to the largest change on the model

prediction significantly increased the adversarial robustness and fairness of deep learning models against a wide range of bias.

Table 1: Performance comparison on the two datasets. MSE of different models on stock volatility prediction n days following the earnings call.

Methods	ACL19			CIKM20		
	MSE ₃	MSE ₇	MSE ₁₅	MSE ₃	MSE ₇	MSE ₁₅
bc-LSTM	1.41	0.44	0.30	1.44	0.51	0.35
MDRM	1.37	0.42	0.30	1.43	0.48	0.32
HTML	0.85	0.35	0.25	1.15	0.47	0.30
M3ANet	0.72	0.38	0.24	0.86	0.41	0.31
AMA-LSTM	0.68	0.36	0.23	0.74	0.35	0.27

Table 2: Δ MSE is the difference in MSE between female and male CEO distributions over 3, 7, and 15 days. Our method shows the best performance.

Methods	ACL19			CIKM20		
	Δ MSE ₃	Δ MSE ₇	Δ MSE ₁₅	Δ MSE ₃	Δ MSE ₇	Δ MSE ₁₅
bc-LSTM	0.38	0.23	0.27	0.42	0.31	0.24
MDRM	0.30	0.11	0.28	0.36	0.33	0.29
HTML	0.33	0.14	0.28	0.29	0.25	0.16
M3ANet	0.24	0.10	0.26	0.31	0.28	0.20
AMA-LSTM	0.19	0.08	0.15	0.23	0.21	0.13

5 Experiments

We conducted experiments to answer the following questions about the performance of AMA-LSTM: Q1. **Robustness.** Since earnings calls often contain much extraneous information unrelated to company performance, whether the adversarial training reduce stochastic of the financial data.

Q2. **Fairness.** Due to the differences in verbal and vocal cues across genders, predictions made using gender-specific data may yield varied results. The question arises whether adversarial training can reduce the bias originating from these differences.

Q3. **Ablation study.** Between text and audio information, which exhibits greater bias.

5.1 Dataset

We utilized two real-world datasets from the financial industry: ACL19 (Qin and Yang, 2019) contains 576 public earnings calls audio recordings with their transcripts for 277 companies in the S&P 500. CIKM20 (Li et al., 2020) comprise 3443 earnings call embeddings along with transcripts for 1213 companies in the S&P 1500. Additionally, we sourced the necessary dividend-adjusted closing prices for volatility prediction via the Yahoo Finance³ API by using the relevant stock tickers. Besides, due to various corporate changes such as

³<https://finance.yahoo.com/>

mergers, acquisitions, and rebranding, several companies have changed their names and stock tickers. We manually collected information regarding these stocks from Investing⁴. In terms of data processing, we followed the approach outlined by (Sawhney et al., 2021a) for textual features, employing FinBERT embeddings with their default settings⁵(Araci, 2019). For audio cues, we extracted multi-dimensional feature vectors using Praat (Boersma and Van Heuven, 2001).

5.2 Evaluation metrics

Following (Qin and Yang, 2019; Sawhney et al., 2021a), we evaluate the accuracy of our volatility predictions by comparing the predicted values y_i with the actual volatility values \hat{y}_i . We use the Mean Squared Error (MSE) as our evaluation metric, defined as:

$$MSE = \frac{\sum_i (y_i - \hat{y}_i)^2}{n}. \quad (6)$$

Furthermore, to evaluate gender bias in our model, we measure the disparity in performance errors between genders, denoted as $\Delta MSE = MSE_f - MSE_m$, where f represents female m represents male. A higher value of ΔMSE indicates a bias favoring male-oriented data.

5.3 Baselines

In this section, we compare four baselines. These methods represent previous approaches to stock volatility prediction that utilized either LSTM or transformer-based multimodal models.

- bc-LSTM** (Poria et al., 2017) : employs separate contextual Bi-LSTMs to extract uni-modal features then fused together.

- MDRM** (Qin and Yang, 2019) : utilizes pretrained GloVe embeddings and hand-crafted acoustic features. These are processed through separate BiLSTMs to obtain their uni-modal contextual embeddings. Then, these embeddings are fused and input into a two-layer dense network.

- HTML** (Yang et al., 2020) : employs WWM-BERT to encode text tokens. It then fuses the unimodal features and inputs them into a sentence-level transformer.

- M3ANet** (Sawhney et al., 2021b) : utilizes uncased base BERT to encode text tokens. It attentively fuses the unimodal features and then inputs them into a sentence-level transformer.

⁴<https://www.investing.com/>

⁵<https://github.com/ProsusAI/finBERT>

Table 3: Performance of stochastic perturbation on the two datasets

Datasets	MSE ₃	MSE ₇	MSE ₁₅
ACL19	0.81	0.38	0.28
CIKM20	0.97	0.48	0.32

Performance Comparison. Table 1 displays the prediction performance, measured by MSE, of the compared methods on two datasets. And Table 2 displays the differences in prediction accuracy between earnings calls led by female and male CEOs. The result leads to the following observations:

- Our method achieves the best MSE in almost all cases. It surpasses the previous state-of-the-art Transformer-based methods and LSTM-based approaches. Specifically, AMA-LSTM beats Transformer-based M3ANet by 20.83% and 13.95% across two datasets. These results indicate that through simulated perturbations during adversarial training, the model creates areas resistant to randomness information around the input space. This reduces the interference of stochasticity in financial data on the model’s predictions, thereby enhancing the model’s robustness.
- We evaluate the gender bias by comparing Δ MSE, showing that AMA-LSTM achieves the best results in both short-term and long-term across two datasets compared to models without adversarial training. These results demonstrate that adversarial training diminishes the model’s sensitivity to gender-specific features, thereby improving the fairness. It also highlights the need for increased attention to bias issues within financial data.

Stochastic perturbation. We also show the robustness of adversarial training by comparing its effectiveness against both adversarial and random perturbations. Stochastic multimodal attentive LSTM (SMA-LSTM) is a variation of AMA-LSTM. It creates additional examples by introducing random perturbations to clean input examples. Table 3 displays SMA-LSTM’s performance across two datasets. A comparison with Table 3 reveals that: 1) AMA-LSTM significantly outperforms random perturbation. Specifically, adversarial training outperforms stochastic perturbation by 16.05% on the ACL19 and 23.71% on the CIKM. This shows that adversarial perturbations can improve stock

Table 4: An ablation study was conducted on AMA-LSTM. In this context, AMA-LSTM(A) represents the model using only audio, AMA-LSTM(T) stands for model using only text, and the final variant incorporates both audio and text.

Methods	ACL19			CIKM20		
	Δ MSE ₃	Δ MSE ₇	Δ MSE ₁₅	Δ MSE ₃	Δ MSE ₇	Δ MSE ₁₅
AMA-LSTM (A)	0.25	0.14	0.23	0.27	0.26	0.17
AMA-LSTM (T)	0.22	0.13	0.23	0.24	0.23	0.14
AMA-LSTM	0.19	0.08	0.15	0.23	0.21	0.13

volatility prediction by enhancing model robustness. 2) SMA-LSTM exceeds the performance of the LSTM baseline, which is trained only on clean examples. This emphasizes the importance of addressing the stochastic nature of financial data.

5.4 Ablation Study

An ablation study we have constructed two variations alongside the fully-loaded model. Each variant is specifically tailored to handle certain types of input data: AMA-LSTM(A) solely relies on audio data, AMA-LSTM(T) exclusively processes text embedding. As shown in Table 4, we find that audio data contains more bias than textual information. This can be attributed to audio features differ greatly between males and females, and the uneven gender distribution among speakers in earnings calls amplifies this discrepancy. This result indicates that when processing financial data, there should be a greater focus on the bias in audio data. Furthermore, by merging data from various modalities, the model shows reduced sensitivity to bias and enhanced robustness compared to relying on just one modality.

6 Conclusion

Our research has shown that neural networks used for predicting stock market volatility often face challenges in robustness, particularly in handling the stochasticity and bias inherent in financial audio data. To tackle this, we introduced an innovative solution: the adversarial multimodal attentive LSTM. This method employs adversarial training to more effectively simulate the market noise and biases during model training. The experiments on two benchmark datasets not only validated the effectiveness of our approach but also underscored the importance of considering the stochasticity and bias in financial data for stock volatility prediction tasks. The results further revealed that adversarial training significantly enhances the robustness and fairness of the prediction models.

References

- Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Anish Athalye and Nicholas Carlini. 2018. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*.
- Lyda Bigelow, Leif Lundmark, Judi McLean Parks, and Robert Wuebker. 2014. Skirting the issues: Experimental evidence of gender bias in ipo prospectus evaluations. *Journal of Management*, 40(6):1732–1759.
- Benjamin M Blau, Jared R DeLisle, and S McKay Price. 2015. Do sophisticated investors interpret earnings conference call tone differently than investors at large? evidence from short sales. *Journal of Corporate Finance*, 31:203–219.
- Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glott International*, 5(9/10):341–347.
- Carlyn Burris, Hourii K Vorperian, Marios Fourakis, Ray D Kent, and Daniel M Bolt. 2014. Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements.
- Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE.
- Joseph Comprix, Kerstin Lopatta, and Sebastian A Tideman. 2022. The role of gender in the aggressive questioning of ceos during earnings conference calls. *The Accounting Review*, 97(7):79–107.
- Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar. 2021. Fairness measures for machine learning in finance. *The Journal of Financial Data Science*.
- Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.
- Bernard Dumas, Alexander Kurshev, and Raman Uppal. 2009. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance*, 64(2):579–629.
- Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. [Enhancing stock movement prediction with adversarial training](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5843–5849. International Joint Conferences on Artificial Intelligence Organization.
- Karyn Fish, Kathrin Rothermich, and Marc D Pell. 2017. The sound of (in) sincerity. *Journal of Pragmatics*, 121:147–161.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. 2015. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*.
- Sheikh Mohammad Idrees, M Afshar Alam, and Parul Agarwal. 2019. A prediction approach for stock market volatility based on time series data. *IEEE Access*, 7:17287–17298.
- Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.
- Werner Kristjanpoller, Anton Fadic, and Marcel C Minutolo. 2014. Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, 41(5):2437–2442.
- Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070.
- Yifan Liu, Zengchang Qin, Pengyu Li, and Tao Wan. 2017. Stock volatility prediction using recurrent neural networks with sentiment analysis. In *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I 30*, pages 192–201. Springer.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang. 2022. Diverse audio captioning via adversarial training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8882–8886. IEEE.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

- Thien Hai Nguyen, Kiyooki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611.
- Eszter Novák-Tót, Oliver Niebuhr, and Aoju Chen. 2017. A gender bias in the acoustic-melodic features of charismatic speech? In *Proceedings of the International Conference on Spoken Language Processing*, pages 2248–2252. International Speech Communication Association.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.
- Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.
- Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978*.
- Ramit Sawhney, Arshiya Aggarwal, and Rajiv Shah. 2021a. 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*, pages 3751–3757.
- Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Shah. 2021b. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762.
- Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 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)*, pages 8001–8013.
- Uri Shaham, Yutaro Yamada, and Sahand Negahban. 2018. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204.
- Sining Sun, Ching-Feng Yeh, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie. 2018. Training augmentation with adversarial examples for robust speech recognition. *arXiv preprint arXiv:1806.02782*.
- Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8).
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Dan Wang, Tianrui Wang, and Ionuț Florescu. 2020. Is image encoding beneficial for deep learning in finance? *IEEE Internet of Things Journal*, 9(8):5617–5628.
- Shengkun Wang, YangXiao Bai, Taoran Ji, Kaiqun Fu, Linhan Wang, and Chang-Tien Lu. 2023. Stock movement and volatility prediction from tweets, macroeconomic factors and historical prices. In *2023 IEEE International Conference on Big Data (Big-Data)*, pages 1863–1872. IEEE.
- Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Oluwasanmi Koyejo. 2022. **A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–599, Seattle, United States. Association for Computational Linguistics.
- Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Htl: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.
- Linyi Yang, Zheng Zhang, Su Xiong, Lirui Wei, James Ng, Lina Xu, and Ruihai Dong. 2018. Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 441–445. IEEE.
- Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. 2021. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. *arXiv preprint arXiv:2110.15317*.
- Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*.
- Weiguo Zhang, Xue Gong, Chao Wang, and Xin Ye. 2021. Predicting stock market volatility based on textual sentiment: A nonlinear analysis. *Journal of Forecasting*, 40(8):1479–1500.
- Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxiang Fang, and S Yu Philip. 2018. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, 143:236–247.