An Empirical Investigation of Bias in the Multimodal Analysis of Financial Earnings Calls

Ramit Sawhney
IIIT Delhi
ramits@iiitd.ac.in

Arshiya Aggarwal
Adobe India
arsaggar@adobe.com

Rajiv Ratn Shah
IIIT Delhi
rajivratn@iiitd.ac.in

Abstract

Volatility prediction is complex due to the stock market's stochastic nature. Existing research focuses on the textual elements of financial disclosures like earnings calls transcripts to forecast stock volatility and risk, but ignores the rich acoustic features in the company executives' speech. Recently, new multimodal approaches that leverage the verbal and vocal cues of speakers in financial disclosures significantly outperform previous state-of-the-art approaches demonstrating the benefits of multimodality and speech. However, the financial realm is still plagued with a severe underrepresentation of various communities spanning diverse demographics, gender, and native speech. While multimodal models are better risk forecasters, it is imperative to also investigate the potential bias that these models may learn from the speech signals of company executives. In this work, we present the first study to discover the gender bias in multimodal volatility prediction due to gender-sensitive audio features and fewer female executives in earnings calls of one of the world's biggest stock indexes, the S&P 500 index. We quantitatively analyze bias as error disparity and investigate the sources of this bias. Our results suggest that multimodal neural financial models accentuate gender-based stereotypes.[1]

1 Introduction

Earnings calls are publicly available, quarterly conference calls where CEOs discuss their company's performance and future prospects with outside analysts and investors (Qin and Yang, 2019; Sawhney et al., 2020b). They consist of two sections: a prepared delivery of performance statistics, analysis and future expectations, and a spontaneous question-answer session to seek additional information not disclosed before (Keith and Stent, 2019). Researchers have studied the Post Earnings Announcement Drift (PEAD) to observe that statements made by upper management affect the way information is digested and acted upon impacting short-term price movements (Ball and Brown, 1968; Bernard and Thomas, 1989; Yang et al., 2020).

Audio features contextualize text and connotate speaker's emotional and psychological state (Fish et al., 2017; Jiang and Pell, 2017; Burgoon et al., 2015; Bachorowski, 1999). Hence, when used with textual features, audio features significantly determine the effect of earning calls on the stock market (Qin and Yang, 2019; Yang et al., 2020). Past research has shown that audio features such as speakers' pitch, intensity, etc. vary greatly across genders (Mendoza et al., 1996; Burris et al., 2014; Latinus and Taylor, 2012). Moreover, female executives are highly underrepresented in these earnings calls(Agarwal, 2019; Investments, 2017). The variation in audio features is amplified by deep learning models due to a dearth of female training examples and is manifested as a gender bias. The system learns unneeded correlations between stock volatility and sensitive attributes like gender, accent, etc. It further perpetuates gender-based stereotypes and generalizations like female executives are less confident than male executives (Lonkani, 2019), men are assessed as more charismatic than female executives under identical conditions (Novák-Tót et al., 2017), and nurses are female and doctors are male (Saunders and Byrne, 2020). Biased models further perpetuate stereotypes that can harm underrepresented communities, specifically in the financial and corporate world. Novák-Tót et al. (2017) even show that female speakers have to deliver better acoustic-melodic performance to seem as charismatic as men.

Taking a step towards fair risk forecasting models, we analyze gender bias by studying the error disparity in the state-of-the-art for multimodal

[1]Code & Implementation: https://github.com/midas-research/multimodal-bias-naacl

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3751–3757
June 6–11, 2021.

volatility prediction, MDRM (Qin and Yang, 2019).

2 Background: Why Study Bias?

Bias in Finance Public financial data is impacting virtually every aspect of investment decision making (Perić et al., 2016; Brynjolfsson et al., 2011). Prior research shows that NLP methods leveraging social media (Sawhney et al., 2020a), news (Du and Tanaka-Ishii, 2020), and earning calls (Wang and Hua, 2014) can accurately forecast financial risk. Companies and investors use statistical and neural models on multimodal financial data to forecast volatility (Cornett and Saunders, 2003; Trippi and Turban, 1992) and minimize risk. These models although effective, may be tainted by bias due to individual and societal differences, often unintended (Mehrabi et al., 2019). For example, models trained on the audio features extracted from CEO's speech in earnings calls (Qin and Yang, 2019), may be prone to bias given the underrepresentation of several demographics across race, gender, native language, etc. in the financial realm.

Bias in AI Bias is prevalent in AI based neural models owing to the lack of diversity in training data (Torralba and Efros, 2011; Tommasi et al., 2017). The design and utilization of AI models trained on gender imbalanced data, pose potential deprivation of opportunities to underrepresented groups such as females(Niethammer, 2020; Dastin, 2018). With over 75% of AI professionals being men, male experiences also dominate algorithmic creation (Forum, 2018). In terms of natural language representation, embeddings such as word2vec and GloVe, trained on news articles may inherit gender stereotypes (Packer et al., 2018; Bolukbasi et al., 2016; Park et al., 2018). Recent studies also show the presence of bias in speech emotion recognition (Li et al., 2019).

Bias in AI and Finance With the advent of AI and Big Data, companies are intelligently using data to measure performance (Newman, 2020). But seldom do enterprises check on the imbalance in gathered data. Women still represent fewer than 20% positions in the financial-services C-suite (Chin et al., 2018) and only 5% of Fortune-500 CEOs are women (Suresh and Guttag, 2019). Studies show that models trained on gender imbalanced data reduce the chances for women to get capital investments or loans (Gürdeniz et al., 2020). Apart from that, using feature representations intrinsic to different genders can inculcate semantic gender bias (Li et al., 2019; Suresh and Guttag, 2019). Professional studies have found that men tend to self-reference using 'I', 'me' and 'mine' whereas women tend to reference the team, like 'we', 'our' and 'us' (Investments, 2017). Although there is great progress in mitigating bias in text, understanding its presence in multimodal speech based analysis, particularly in real world scenarios like corporate earnings calls analysis remain an understudied yet promising research direction. Another study found that despite having identical credibility, female CEOs are perceived as less capable to attract growth capital (Bigelow et al., 2014).

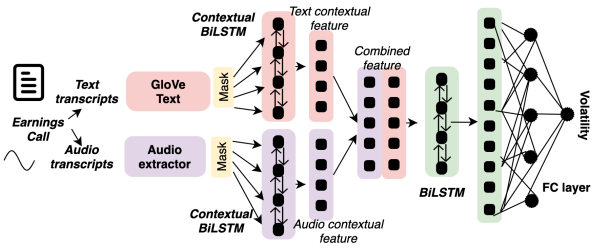

Figure 1: Model architecture used for training the multimodal audio-text model for evaluating the gender specific performance inspired by (Qin and Yang, 2019)

3 Formulation and Experiments

Stock volatility Following Kogan et al. (2009); Sawhney et al. (2020c), for a given stock, with a close price of p_i on trading day i, we calculate the average log volatility over n days following the day of the earnings call as:

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

where, the return price r_i is defined as $\frac{p_i}{p_{i-1}} - 1$ and $\bar{r}$ is the average of r_i from 0 to τ.

Volatility Prediction Consider each earnings call E, with aligned audio recordings A and text transcripts T. The earnings calls are divided into separate distributions based on the gender of the speaker to analyse the effect of gender on the model performance. Building upon the work of Qin and Yang (2019); Yang et al. (2020) our main focus is to learn a function $f(E_{\{T,A\}}) \rightarrow v_{[0,\tau]}$, over $\tau \in \{3, 7, 15, 30\}$ days to evaluate the bias for different time periods.

Split	Date Range (2017)	Female	Male	Total
Train	17 Jan- 3 Aug	11%	89%	391
Val	3 Aug- 24 Oct	12.5%	87.5%	56
Test	24 Oct- 21 Dec	14.3%	85.7%	112
		12%	88%	559

Table 1: Details of the Train, Validation and Test sets

	$\Delta G = MSE_F - MSE_M \downarrow$			
	$\tau = 3$	$\tau = 7$	$\tau = 15$	$\tau = 30$
MDRM(A)	0.38	0.16	0.26	0.18
MDRM(T)	0.33	0.12	0.20	0.16
MDRM(AT)	0.30	0.11	0.28	0.14

Table 2: Modality specific ΔG i.e. the difference between the MSE for female and male distributions for 3, 7, 15 and 30 days over 5 runs. Here A stands for Audio only, T for Text only and AT for Audio and Text.

	Test MSE ↓		
	Combined	**Male**	**Female**
Audio			
$\tau = 3$	$0.738_{\pm 0.03}$	$0.684_{\pm 0.02}$	$1.059_{\pm 0.04}$
$\tau = 7$	$0.395_{\pm 0.03}$	$0.372_{\pm 0.02}$	$0.536_{\pm 0.07}$
$\tau = 15$	$0.292_{\pm 0.02}$	$0.255_{\pm 0.02}$	$0.511_{\pm 0.05}$
$\tau = 30$	$0.208_{\pm 0.02}$	$0.182_{\pm 0.02}$	$0.362_{\pm 0.05}$
Text			
$\tau = 3$	$0.662_{\pm 0.05}$	$0.615_{\pm 0.04}$	$0.943_{\pm 0.09}$
$\tau = 7$	$0.390_{\pm 0.08}$	$0.372_{\pm 0.08}$	$0.495_{\pm 0.11}$
$\tau = 15$	$0.252_{\pm 0.04}$	$0.224_{\pm 0.05}$	$0.419_{\pm 0.08}$
$\tau = 30$	$0.225_{\pm 0.07}$	$0.202_{\pm 0.06}$	$0.362_{\pm 0.10}$
Audio + Text			
$\tau = 3$	$0.644_{\pm 0.08}$	$0.603_{\pm 0.07}$	$0.898_{\pm 0.10}$
$\tau = 7$	$0.362_{\pm 0.08}$	$0.345_{\pm 0.06}$	$0.457_{\pm 0.07}$
$\tau = 15$	$0.308_{\pm 0.07}$	$0.272_{\pm 0.06}$	$0.552_{\pm 0.14}$
$\tau = 30$	$0.185_{\pm 0.02}$	$0.165_{\pm 0.02}$	$0.308_{\pm 0.04}$

Table 3: Test MSE results over 5 runs for the individual and combined Audio-Text modalities and male-female distributions for all time periods i.e. 3, 7, 15, 30 days.

Earnings Call Data We use the dataset[2] created by Qin and Yang (2019) comprising 559 public earnings calls audio recordings with their transcripts for 277 companies in the S&P 500 index spanning over a year of earnings calls. The details of the dataset splits for training have been given in Table 1. For the identification of gender bias in the earnings calls acoustics, we first map the speakers from all the earnings calls to their **self reported** gender. For this we perform web scrapping from Reuters[3] (pronouns), Crunchbase[4] where the genders are self-declared and the available genders from the Wikidata API. The genders extracted correspond only to male and female, 11.8% of the speakers are female and 88.2% are male which motivates us to estimate the error disparity in model performance.

Evaluating Gender Bias We use performance error disparity $\Delta G = MSE_f - MSE_m$ where f and m stand for female and male respectively (Saunders and Byrne, 2020). A higher ΔG is indicative of bias is in favour of the male distribution.

Model Architecture and Training We use the state-of-the-art, Multimodal Deep Regression Model (MDRM) Qin and Yang (2019), as shown in Figure 1. MDRM takes utterancde level audio A and text T embeddings and models them through two contextual BiLSTM layers followed by late multimodal fusion. The fused text-audio features are fed to another BiLSTM followed by two fully-connected layers. MDRM is trained end-to-end by optimizing the mean square error (MSE) between the predicted and true stock volatility.

Training Setup For textual features we use FinBERT embeddings[5] (Araci, 2019) with default parameters and for audio cues, we extract 26-dimensional vectors with Praat (Boersma and Van Heuven, 2001) extracted by Qin and Yang (2019), spanning Shimmer, Jitter, Pitch, Intensity, etc. We report the complete list in Table 4. The maximum number of audio clips in any call is 520. Hence, we zero-pad the calls that have less than 520 clips. The model is trained on TPU version 3.8 for 20 epochs using a learning rate of 0.001. The hyperparameters are tuned on the validation set defined by Qin and Yang (2019) following the same preprocessing. We perform 5 end-to-end runs with early stopping over the validation loss to arrive at the decision of training for 20 epochs.

4 Results and Analysis

Bias in Multimodal Volatility Prediction For evaluating gender bias in MDRM, we analyze the error disparity quantified by ΔG for the individual text and audio modalities and their combination for τ = 3,7,15,30 days. We tabulate the error disparity

[2] https://github.com/GeminiLn/EarningsCall_Dataset

[3] https://www.thomsonreuters.com/en/profiles.html

[4] https://www.crunchbase.com/discover/people

[5] https://github.com/ProsusAI/finBERT

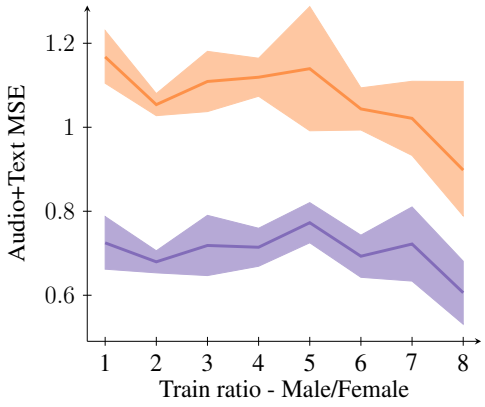

Figure 2: MSE mean values for 5 runs of different train set distributions for multimodal MDRM with $\tau = 3$. Shading indicates error bounds upto the first standard deviation across MSE for 5 independent runs per training ratio split.

in terms of ΔG across modalities in Table 2 and performance in Table 3. We observe that for all modalities the error for male distribution is consistently less than that of female distribution for both short- and long-term durations. Although the audio modality improve model performance significantly, it has the highest amount of bias as audio features for males and females vary significantly. Further, the skewed distribution of speakers' gender in the earnings calls amplifies this error disparity.

Over amplification refers to bias that occurs in a system during model fitting. The model learns imperfect generalizations between the attributes and the final labels and amplifies them while predicting on the test set. In our case, since female examples are very less in comparison to the male counterparts, the model discriminates between male and female examples by inferring insufficient information beyond its source base rate as shown in Table 2. To study this effect we train the model for different training sample ratios as per gender to observe performance variation in Figure 2. We note that as the male:female training ratio increases, the test loss is amplified the most in audio modality followed by audio+text and text. Test MSE_{male} decreases in comparison to increase in MSE_{female}. MSE_{female} increases as the percentage of female examples in the train set decreases as the generalised notions of this underrepresented community are learnt and the incorrect inferences do not harm the overall performance much. Since the difference in test loss for male and female is significantly less when the number of samples across genders is equal. Through this observation, we note that performance for female examples can be improved by augmentation techniques or cross domain adaptation, which we leave for future work.

Audio features	***P* value**	**Bonferroni**
Pitch Analysis		
Mean Fundamental Frequency (F0)	↑	
Stdev Fundamental Frequency (F0)	↑↑↑	
Number of pulses	↓↓↓↓	*
Number of periods	↓↓↓↓	*
Degree of voice breaks	↑↑↑	
Maximum Pitch	↓	
Minimum Pitch	↓	
Voiced Frames	↑	
Voiced to Unvoiced Ratio	↑	
Voiced to Total Ratio	↑	
Intensity Analysis		
Mean Intensity	↑↑	
SD Energy	↑↑↑	
Maximum Intensity	↑↑	
Minimum Intensity	↓	
Voice Analysis		
Local Jitter	↑↑↑↑	*
Local Absolute Jitter	↑↑↑↑	*
Relative Average Perturbation Jitter	↑↑↑↑	*
Period Perturbation Quotient-5 Jitter	↑↑↑↑	*
ddp Jitter	↑↑↑↑	*
Local Shimmer	↑↑↑↑	*
Local dB Shimmer	↑↑↑	*
apq3 Shimmer	↑↑↑↑	*
apq5 Shimmer	↑↑↑↑	*
apq11 Shimmer	↑↑↑↑	*
dda Shimmer	↑↑↑↑	*
Harmonicity Analysis		
Harmonic to Noise Ratio	↓↓↓	

Table 4: Comparison of the audio features for male and female speaker distributions. The number of bars signify the magnitude of the P-value and the direction indicates the relation of the mean of the male distribution with that of the female distribution. ↑↑↑↑ : mean of male is higher than female with $P < 0.001$, ↑↑↑ : $P < 0.01$, ↑↑ : $P < 0.05$, ↑ : $P >= 0.05$. Features whose difference is statistically significant for the male and female distributions under the two-tailed T-test after the Bonferroni correction are marked with *.

Semantic Bias occurs in embeddings and representations of audio and textual data which learn unwanted stereotypes. For our case semantic bias occurs as the audio features are significantly different for male and female distributions. We analyze each audio feature for both distributions in Table 4. We find that 13 out of 26 features have a statistically significant difference under the two-tailed T-test ($\alpha = 0.05$) after applying Bonferroni correction (Weisstein, 2004), a multiple comparison

correction when multiple statistical tests are being performed. These differences in audio features of executives' speech can amplify the error disparity, as models may associate certain gender specific features such as Voice analysis-based features like Shimmer and Jitter.

5 Ethical Considerations

Degradation in the performance of speech models could be due to discernible noise and indiscernible sources like demographic bias: age, gender, dialect, culture, etc (Meyer et al., 2020; Hashimoto et al., 2018; Tatman and Kasten, 2017). Studies also show that AI can deploy biases against black people in criminal sentencing (Angwin et al., 2016; Tatman and Kasten, 2017). Although we only account for the gender bias in our study, we acknowledge that there could exist other kinds of bias due to age, accent, culture, ethnic and regional disparities in audio cues, as the publicly available earnings calls majorly have companies belonging to the US. Moreover, only publicly available earnings calls have been used limiting the scope of the data. This also limits the availability of genders in the data to only male and female. In the future, we hope to increase the amount of data to expand our study to more categories and types of sensitive attributes.

6 Conclusion

Earnings calls provide company insights from executives proving to be high risk-reward opportunities for investors. Recent multimodal approaches that utilize these acoustic and textual features to predict the financial risk achieve state-of-the-art performance, but overlook the gender bias associated with speech. We analyze the gender bias in volatility prediction of earnings calls due to gender sensitive audio features and underrepresentation of women in executive positions. We observe that the while adding speech features improves performance, it also perpetuates gender bias, as the audio modality has the highest error disparity. We further probe into the sources of bias, and analyze audio feature variations across gender, and perform experiments with varying training data distributions. Our study presents the first analysis of its kind to analyze gender bias in multimodal financial forecasting to bridge the gap between fairness in AI, neural financial forecasting and multimodality.

References

Sray Agarwal. 2019. Fair AI: How to Detect and Remove Bias from Financial Services AI Models. [Online; accessed 11-September-2019].

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. [Online; accessed 23-May-2016].

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Jo-Anne Bachorowski. 1999. Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57.

Ray Ball and Philip Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of accounting research*, pages 159–178.

Victor L Bernard and Jacob K Thomas. 1989. Post-earnings-announcement drift: delayed price response or risk premium? *Journal of Accounting research*, 27:1–36.

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.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot Int*, 5:341–347.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Erik Brynjolfsson, Lorin M Hitt, and Heekyung Hellen Kim. 2011. Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Judee Burgoon, W. Mayew, Justin Giboney, Aaron Elkins, Kevin Moffitt, Bradley Dorn, Michael Byrd, and Lee Spitzley. 2015. Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls. *Journal of Language and Social Psychology*, 35.

Carlyn Burris, Houri K Vorperian, Marios Fourakis, Ray D Kent, and Daniel M Bolt. 2014. Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research*.

Stacey Chin, Alexis Krivkovich, and Marie-Claude Nadeau. 2018. Closing the gap: Leadership perspectives on promoting women in financial services. [Online; accessed 06-September-2018].

Marcia Millon Cornett and Anthony Saunders. 2003. *Financial institutions management: A risk management approach*. McGraw-Hill/Irwin.

Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. [Online; accessed 10-October-2018].

Xin Du and Kumiko Tanaka-Ishii. 2020. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3353–3363, Online. Association for Computational Linguistics.

Karyn Fish, Kathrin Rothermich, and Marc D Pell. 2017. The sound of (in) sincerity. *Journal of Pragmatics*, 121:147–161.

World Economic Forum. 2018. Global Gender Gap Report 2018 - Assessing Gender Gaps in Artificial Intelligence.

Ege Gürdeniz, Elizabeth St-Onge, and Madeline Kreher. 2020. How Artificial Intelligence Can Perpetuate Gender Imbalance. [Online; accessed March-2020].

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*.

American Century Investments. 2017. Evidence of Gender Bias: An Analysis of Conference Call Dialogue. [Online; accessed 09-November-2017].

Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication*, 88:106–126.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

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, Boulder, Colorado. Association for Computational Linguistics.

Marianne Latinus and Margot J Taylor. 2012. Discriminating male and female voices: differentiating pitch and gender. *Brain topography*, 25(2):194–204.

Zhixuan Li, Liang He, Jingyang Li, Li Wang, and Wei-Qiang Zhang. 2019. Towards discriminative representations and unbiased predictions: Class-specific angular softmax for speech emotion recognition. In *INTERSPEECH*, pages 1696–1700.

Ravi Lonkani. 2019. Gender differences and managerial earnings forecast bias: Are female executives less overconfident than male executives? *Emerging Markets Review*, 38:18–34.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Elvira Mendoza, Nieves Valencia, Juana Muñoz, and Humberto Trujillo. 1996. Differences in voice quality between men and women: use of the long-term average spectrum (ltas). *Journal of voice*, 10(1):59–66.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6462–6468.

Daniel Newman. 2020. Why The Future Of Data Analytics Is Prescriptive Analytics. [Online; accessed 02-January-2020].

Carmen Niethammer. 2020. AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators. [Online; accessed 02-March-2020].

E Novák-Tót, O Niebuhr, and A Chen. 2017. A gender bias in the acoustic-melodic features of charismatic speech?(pp. 2248–2252). stockholm, sweden: Proc. In *18th International Interspeech Conference*.

Ben Packer, Yoni Halpern, Mario Guajardo-Céspedes, and Margaret Mitchell. 2018. Text Embedding Models Contain Bias. Here's Why That Matters. [Online; accessed 13-April-2018].

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Amela Perić, Nedžad Polic, and Emira Kozarevic. 2016. Application of data science in finance and other industries.

Yu Qin and Yi Yang. 2019. 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)*, page 390.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020b. 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, Online. Association for Computational Linguistics.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020c. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 456–465, New York, NY, USA. Association for Computing Machinery.

Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, pages 934–938.

Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Robert R Trippi and Efraim Turban. 1992. *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill, Inc.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1155–1165.

Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.