Volume 11, Issue 3, March – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

A Comprehensive Review of Multimodal Financial Sentiment Analysis

Jin Wang[1]; Inam Ullah[2]*

[1,2]Faculty of Information Technology, City University Malaysia, Petaling Jaya 46100, Selangor, Malaysia
[1]ORCID:https://orcid.org/0009-0006-3105-8434
[2]ORCID: https://orcid.org/0009-0005-9298-053X

Corresponding Author: Inam Ullah*

Publication Date: 2026/03/30

Abstract: This study aims to explain how textual content, vocal cues, and affective expressions jointly shape investor reactions and market fluctuations by synthesizing existing research on multi-modal sentiment analysis in financial settings. It uses earnings conference calls as a representative example. It adopts a structured literature review approach, organizing and comparing prior work across theoretical foundations, data and feature construction, model architectures, and fusion strategies, including domain-specific language models and multi-modal Transformer frameworks. The review concludes that multi-modal methods generally outperform text-only approaches because acoustic signals capture soft information, such as managerial uncertainty, stress, and confidence, thereby improving the modelling of market reactions and return-related outcomes. However, progress is constrained by scarce and heterogeneous multimodal datasets, imperfect cross-modal temporal alignment, and limited transparency and causal identification, which together hinder reproducibility, generalizability, and real-time deployment in practice.

Keywords: *Multimodal Sentiment Analysis; Earnings Conference Calls; Financial Communication; Vocal Emotion; Explainable Artificial Intelligence; Causal Inference; Behavioral Finance.*

How to Cite: Jin Wang; Inam Ullah (2026) A Comprehensive Review of Multimodal Financial Sentiment Analysis. *International Journal of Innovative Science and Research Technology*, 11(3), 2596-2603. https://doi.org/10.38124/ijisrt/26mar1387

I. INTRODUCTION

When we talk about current and up-to-date financial markets, intangible signals such as language, voice, and emotional tone are increasingly recognized as influential drivers of investor behavior and asset pricing (Anastasiou et al., 2025). As analysts, institutional investors, and media outlets scrutinize not only what executives say but also how they say it, managerial communication has become a key source of soft information, and earnings conference calls are a representative setting where executives discuss financial results and future outlooks in real time, providing a rich combination of textual, vocal, and emotional cues for sentiment analysis. Traditional financial sentiment analysis has largely focused on textual sources such as earnings call transcripts, press releases, and financial reports using lexicon-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

based methods or supervised classifiers, but these approaches capture only part of the communicative signal. Recent advances in machine learning, particularly in natural language processing and speech recognition, enable analysis of both content and delivery, incorporating vocal prosody, emotional tone, and hesitation cues to better reflect managerial intent and investor perception (Gill et al., 2026; Baik et al., 2024). Integrating multimodal data, including text and audio and facial expression in some cases, has further improved predictive performance, as vocal cues such as pitch, rhythm, and emotional congruence can signal uncertainty, confidence, or deception and may precede market reactions (Chen et al., 2025; Ewertz et al., 2025), with evidence that a CEO's hesitant tone in forward-looking statements can be associated with negative abnormal returns even when verbal content appears neutral and that emotionally charged language can trigger stronger investor responses than purely factual disclosures (Ewertz et al., 2025). Despite this momentum, the field remains fragmented and methodologically heterogeneous, with limited consensus on model architectures, fusion strategies, and evaluation metrics, while explainability and causal inference are still insufficiently integrated into most multimodal frameworks, leaving open questions about interpretability, generalizability, and real-world usability. To address these gaps, this study uses a structured literature review approach to synthesize prior work on data sources, feature extraction, model architectures, multimodal fusion, explainable AI, and causal inference in earnings-call-based financial sentiment analysis.

II. BACKGROUND AND PROBLEM CONTEXT

This section establishes the financial phenomenon and the communication setting that motivate multimodal financial sentiment research. It highlights that delayed market reactions after earnings news and the rich, multimodal nature of earnings calls jointly create both a practical need and an analytical opportunity for multimodal modeling.

➢ *Post-Earnings Announcement Drift and Market Implications*

The post-earnings announcement drift (PEAD) remains one of the most persistent and well-documented anomalies in financial economics. It refers to the tendency of a firm's stock price to continue moving in the direction of an earnings surprise, positive or negative, for days, weeks, or even months following the announcement. This pattern stands in contrast to the semi-strong form of the Efficient Market Hypothesis (EMH), which asserts that all publicly available information, including earnings releases, should be rapidly and fully reflected in stock prices (Fama, 1970). The continued presence of PEAD, despite decades of empirical scrutiny, raises critical questions about investor cognition, information assimilation, and the limits of market efficiency. The roots of PEAD can be traced to the seminal work of Ball and Brown (1968), who showed that stock prices not only react at the time of earnings announcements but also continue to drift in the direction of the surprise afterward. Their event study methodology paved the way for more systematic investigations. Bernard and Thomas (1989, 1990) provided stronger statistical evidence, revealing significant serial correlation between the magnitude of earnings surprises and post-announcement returns, even after adjusting for firm- and market-level factors. Subsequent research has confirmed the robustness of PEAD across time periods, industries, and geographical markets. Whether measured over short windows or longer horizons, the drift remains both statistically significant and economically relevant. Notably, the magnitude of the drift tends to scale with the size of the earnings surprise and is often more pronounced following positive surprises than negative ones (Livnat & Mendenhall, 2006). Its resilience across different market regimes and regulatory environments suggests that PEAD is not merely the product of temporary inefficiencies.

➢ *Earnings Conference Calls as a Multimodal Disclosure Setting*

Earnings calls have emerged as one of the most information-dense and strategically significant communication platforms between publicly traded firms and their stakeholders. Held on a quarterly basis, these events serve not only to disclose financial results but also to shape narratives, manage market expectations, and signal future strategic direction. Unlike static disclosures such as financial statements or press releases, earnings calls are inherently dynamic and multimodal, encompassing verbal language, vocal tone, emotional inflection, and real-time interaction. This multimodal composition presents both analytical challenges and opportunities, particularly in understanding investor behavior and anomalies such as post-earnings announcement drift (PEAD). A standard earnings call typically

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

comprises two distinct segments, prepared remarks and a question-and-answer (Q&A) session. During the prepared remarks, senior executives, most commonly the Chief Executive Officer (CEO) and Chief Financial Officer (CFO), deliver a scripted overview of the firm's financial performance, recent developments, and forward-looking guidance.

In contrast, the Q&A session introduces an unscripted and interactive format. Analysts pose spontaneous questions, often probing beyond the scope of the prepared statements or challenging inconsistencies, and executives must respond in real time, frequently revealing subtle emotional cues such as hesitation, defensiveness, or assertiveness (Matsumoto, Pronk, & Roelofsen, 2011). This bifurcated structure offers a natural delineation for empirical analysis. Researchers have found that the Q&A segment generally contains greater predictive power for future firm performance than the scripted portion, suggesting that investors place considerable emphasis on spontaneous managerial communication (Mayew & Venkatachalam, 2012; Larcker & Zakolyukina, 2012).

III. REVIEW METHODOLOGY

The study uses a structured literature review method. The review article is organized to mirror the established logic, progressing from the Post-Earnings Announcement Drift (PEAD) anomaly to multimodal disclosures, sentiment modeling, and the integration of causal inference and Explainable AI (XAI) as pillars of trustworthiness to ensure thematic consistency.

➢ *Review Scope and Conceptual Framework*

The scope of this review is anchored in the premise that earnings calls are dynamic and multimodal environments encompassing verbal language, vocal tone, emotional inflection, and real-time interaction. Evidence is synthesized using a coding frame aligned with the following thematic sections:

- Text-based sentiment analysis transitions from traditional lexicons to machine learning and transformer-based models like FinBERT.
- Acoustic and emotional signals include prosody, hesitations, and stress markers analyzed through affective computing frameworks.
- Multimodal fusion and alignment explore early, late, and attention-based strategies alongside synchronization challenges.
- Trustworthy modeling focuses on causal inference techniques, such as Propensity Score Matching (PSM) and Directed Acyclic Graphs (DAG), as well as XAI tools like SHAP and LIME.

➢ *Search Strategy*

The search strategy reflects a dual focus on established finance and accounting evidence and rapidly evolving AI methodologies. The review draws from peer-reviewed journals in finance, information systems, and machine learning, supplemented by technical conference proceedings and preprints where novel multimodal architectures often debut. Search strings were constructed by combining disclosure context terms with multimodal technical descriptors and economic outcome variables, such as market reaction, abnormal returns, and trading volume, to ensure a comprehensive narrative flow from PEAD to advanced modeling.

➢ *Inclusion and Exclusion Criteria*

Studies were selected for inclusion based on three primary criteria:

- The examination of financial communication, specifically earnings calls, as the primary disclosure setting.
- The operationalization of two or more modalities, with text and audio serving as the minimum requirement.
- The presence of a link between multimodal signals and economically meaningful outcomes or validated behavioral interpretations.
- Conversely, studies were excluded if they relied strictly on text-based analysis or lacked sufficient methodological detail regarding feature extraction and fusion, which are identified as critical barriers to reproducibility.

➢ *Screening and Coding Procedure*

The screening process followed a staged approach, beginning with title and abstract reviews followed by full-text assessments to confirm the presence of a genuine multimodal pipeline. Eligible studies were then coded using a standardized extraction template that captures disclosure structure, text representations, and acoustic features such as pitch and MFCCs. Additionally, the coding captured fusion strategies, alignment methods, evaluation designs—including text-only

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

baselines—and the specific XAI or causal identification tools employed.

➢ *Synthesis Approach and Trustworthiness Emphasis*

Given the studyexplicit observation that the literature is constrained by data scarcity, heterogeneous pipelines, and inconsistent reporting, findings are synthesized using a structured narrative-comparative approach rather than meta-analysis.The synthesis prioritizes (1) whether multimodal gains are demonstrated against strong text-only baselines, (2) whether the study reports alignment and preprocessing choices transparently, and (3) whether it advances beyond correlation through causal inference or improves usability via explainability.

IV. MULTIMODAL FINANCIAL SENTIMENT MODELING

This section synthesizes how multimodal financial sentiment models are built and evaluated, focusing on three core components: text-based sentiment extraction, acoustic and affective signals in managerial speech, and multimodal fusion and alignment strategies.

➢ *Text-Based Financial Sentiment Analysis*

Financial sentiment analysis has evolved from early lexicon-based approaches toward machine-learning and deep-learning pipelines that better capture context and domain-specific meaning. Lexicons remain useful for transparency and speed, but they often struggle with ambiguity, negation, and finance-specific language. Supervised learning introduced more adaptable classifiers, while deep learning improved the modeling of sequential and contextual dependencies in financial documents. Recent transformer-based language models further strengthened representation learning for finance, leveraging contextual embeddings and transfer learning, including Bidirectional Encoder Representations from Transformers introduced by Devlin et al. (2019). For example, finance-specific transformer models such as FinSentiment demonstrate that pretraining BERT-, RoBERTa-, GPT-, and LLaMA-based architectures on financial corpora significantly improves sentiment classification performance compared to general-domain models (Ergun & Sefer, 2025). In earnings-related applications, transcripts are a central textual source because they combine prepared remarks and interactive Q&A content, enabling fine-grained sentiment measurement and links to market outcomes. Prior work also emphasizes that financial language is frequently strategic and highly contextual, making robustness across firms, time periods, and reporting regimes a persistent challenge. As a result, text-based modeling in this domain commonly prioritizes domain adaptation, careful evaluation design, and outcome-oriented validation using event-based market indicators, taking earnings call transcripts as a representative dataset (Todd et al. , 2024;Du et al., 2024; Mai et al.,2024)

➢ *Acoustic and Affective Signals in Managerial Speech*

Managerial speech conveys information beyond words, because vocal delivery can reflect affective and cognitive states that shape how investors interpret disclosures. Voice is widely treated as a psychological signal that carries emotion and intention through prosodic variation, including pitch, intensity, rhythm, and pauses, and these cues can influence perceived confidence or uncertainty (Baik et al., 2025; Ewertz, 2026; Gupta, 2025). In financial contexts, acoustic feature extraction typically begins with low-level descriptors and summary statistics, and then moves toward learned representations that aim to reduce noise from recording conditions and speaker heterogeneity. Emotional speech analysis often adopts dimensional emotion representations, including valence and arousal, to connect vocal patterns to interpretable affective states, with the valence–arousal view commonly traced to Russell (1980). Empirical work summarized in the file highlights that these signals are especially salient in unscripted segments such as Q&A, where hesitation and stress-related patterns can emerge more clearly than in prepared remarks. Overall, acoustic and affective modeling complements text by capturing “how it is said,” which can strengthen predictions of market reactions when integrated with textual sentiment.

➢ *Multimodal Fusion and Alignment Strategies*

Multimodal financial sentiment modeling relies on fusion strategies that combine modalities while preserving complementary information and minimizing redundancy. Core approaches include early fusion that learns joint representations from concatenated features, late fusion that aggregates decisions from modality-specific models, and attention-based or hierarchical fusion that dynamically weights modalities by context. The file emphasizes that

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

alignment and synchronization are central technical issues, because text and audio unfold over time and may not correspond at a one-to-one granularity, making segmentation and temporal matching critical. Multimodal learning research also frames fusion as a problem of representation, translation, and alignment across heterogeneous signals, with a commonly cited conceptual overview (Li & Tang, 2024; Baltrušaitis et al., 2019). Evaluation in multimodal settings therefore needs to consider both predictive gains and stability across datasets, speakers, and market periods, because improved accuracy can be undermined by misalignment, noisy audio, or shifting discourse patterns. In practice, robust multimodal pipelines tend to pair fusion with alignment-aware preprocessing and transparent ablation testing to show how much each modality contributes under comparable conditions.

V. TRUSTWORTHY MODELING FOR FINANCE

This section focuses on two pillars that make multimodal financial sentiment models usable in high-stakes settings: explainability, which clarifies why models predict certain market responses, and causal inference, which strengthens claims about whether sentiment signals drive outcomes rather than merely correlate with them.

➢ *Explainable AI for Multimodal Financial Models*

As financial institutions increasingly deploy deep learning systems to analyze managerial communication, investor sentiment, and market dynamics, the demand for model transparency has grown accordingly. Unlike consumer applications where predictive accuracy may suffice, financial decision-making often requires that models be interpretable, auditable, and justifiable, particularly in contexts involving capital allocation, regulatory compliance, and fiduciary obligations. In multimodal financial models—where predictions are derived from the integration of textual, acoustic, and emotional inputs—explainability becomes even more essential, because these models operate on intricate signals in high-stakes environments where credibility, accountability, and theoretical interpretability are indispensable. Explainable AI aims to close the gap between model complexity and human understanding by clarifying how and why a model arrives at specific predictions, and in multimodal sentiment analysis it can reveal which words, vocal patterns, or emotional cues drive decision-making. Common tools include LIME (Ribeiro et al., 2016) , which perturbs inputs near a given prediction and fits a simple interpretable model, and SHAP (Lundberg & Lee, 2017) , which assigns contribution scores to each feature based on cooperative game theory. In financial forecasting tasks that integrate sentiment features, SHAP is often used to attribute the contribution of sentiment-related predictors to model outputs (Ruan & Jiang, 2025). Attention visualization is also widely used, such as in multimodal transformers where cross-attention maps can reveal alignments between textual content and acoustic peaks, although attention is not always a faithful proxy for feature importance (Yamaguchi et al., 2025). A distinct challenge in multimodal modeling is modality attribution, which assesses how much each modality contributes to a prediction, and this supports both stakeholder trust and behavioral insight when models are used for decision support.

➢ *Causal Inference for Sentiment and Market Response*

A foundational framework in causal inference is the Rubin Causal Model, also known as the potential outcomes framework, where each observation has two potential outcomes under treatment and control, and the causal effect is defined as the difference between these hypothetical outcomes, although only one outcome is observed in reality. Propensity Score Matching is widely used to estimate treatment effects from observational data by matching treated and control observations with similar propensity scores, and in financial sentiment it has been employed to isolate the effect of managerial tone. Huang et al. (2021), for instance, matched earnings calls with similar financial fundamentals but differing vocal sentiment, finding that optimistic-sounding calls elicited stronger short-term stock responses, suggesting a causal effect of delivery style independent of content. To mitigate unobserved confounding, instrumental variable methods can be used when an instrument correlates with sentiment tone but affects returns solely through that tone, and identification examples include exogenous variation such as call timing that may influence vocal delivery without directly affecting stock prices. In some frameworks, sentiment is treated as a mediator, and mediation analysis can decompose total effects into direct and indirect components using counterfactual-based approaches (Da et al., 2024; Huang, 2023;Hoekstra & Güler, 2022). Structural Causal Models and Directed Acyclic Graphs provide a formal representation of assumed causal

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

relationships (Da et al., 2024), and Zhang et al. (2022) used this framework to isolate the effect of vocal disfluency on investor reaction while controlling for confounding textual features. Recent work also uses deep generative models to simulate counterfactual earnings calls, and Wang et al. (2023) employed a conditional VAE to generate alternative acoustic profiles to estimate individualized treatment effects of vocal sentiment on stock returns.

VI. CONCLUSION

As markets become faster, more interconnected, and more responsive to soft information, the language and vocal nuances of corporate managers have become important signals to investors, analysts, and regulators. This review has basically brought together all the research on the topic of multimodal sentiment analysis in finance. It is very focused on the combination of signals - text, voice, and emotion - coming from managerial communication. The main point made here is that managerial communication is essentially multimodal. The text part communicates the stories and data, the voice reveals the psychological and emotional states, and the sentiment is the behavioral signal that influences the investor's reaction. We have followed the timeline of sentiment analysis from the very first lexicon-based methods to the latest, more context-sensitive models like FinBERT and multimodal transformers that can be used in parallel with speech processing techniques that allow the extraction of prosodic and affective features from audio. Researchers, by combining these modes via early, late, or attention-based architectures, have been able to significantly outperform traditional models in predicting returns, estimating volatility, and identifying earnings surprises. However, such progress is not without downsides since there are still issues like lack of data, timing alignment between different data, making the model understandable to people, and overfitting in the situations of small data. Also, it has been noticed that implementing multimodal systems in live financial scenarios is not very common due to various difficulties in these environments. Explainability has become a major research direction initiated partly by the need to overcome these issues, and also, causal inference techniques have been introduced to separate behavioral effects from statistical artifacts progressively. This work increases both the ability to understand sentiment models and their trustworthiness. Moreover, the application of causal inference methods has come along the explanation of the models, advancing the interpretability and the credibility of the models. What is more, large-scale multilingual multimodal datasets, real-time and adaptive sentiment systems for live earnings calls, stronger integration of causal frameworks, investor heterogeneity modeling improvements, and human-centered explainability tools supported by deeper interdisciplinary collaboration will guide the future of the field. In summary, multimodal sentiment analysis in finance is not simply an engineering task. It is a question of rethinking how we communicate and understand information within the financial ecosystems, and it is a very promising approach to capturing the richness of financial communication while moving toward models that can predict and explain market behavior.

VII. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

As a structured literature review, this article has several methodological limitations. (1) A literature synthesis is always limited by the scope of the domain surveyed; in this case, evidence is predominately focused on earnings conference calls and reflects mainly English-language disclosures from U.S.-centered markets, thus may have less cross-market generalizability. (2) The literature base is very diverse in the types of tasks, datasets, feature definitions, fusion designs, and evaluation windows, thus the papers are hardly directly comparable and at the same time, the effect sizes cannot be quantitatively meta-analyzed in a trustworthy way. (3) There are very few multimodal studies of uniform reporting practices, as information about preprocessing, alignment procedures, and ablation settings is disclosed inconsistently, thereby lowering reproducibility and weakening cumulative inference. (4) There may be publication and selection bias, since positive predictive gains and novel architectures are more likely to be published than null findings or robustness failures, and thus they become biased. Finally, because multimodal finance is developing so quickly, the review may not include the very latest model releases and newly constructed datasets, which thereby necessitates that the review periodically be updated and standardized benchmarking used. Therefore, future research can prioritize transparent and replicable review protocols, expand coverage to multilingual and non-U.S. markets, and develop standardized task definitions, alignment pipelines, and shared benchmarks so that results across studies

become directly comparable and suitable for meta-analytic synthesis.

REFERENCES

[1]. Anastasiou, D., Katsafados, A., Ongena, S., & Tzomakas, C. (June 19, 2025). *Beyond words: Fed chair voice sentiments and US bank stock price crash risk*. VoxEU/CEPR. https://cepr.org/voxeu/columns/beyond-words-fed-chair-voice-sentiments-and-us-bank-stock-price-crash-risk

[2]. Baik, B.; Kim, A. G.; Kim, D. S.; Yoon, S. (2025). Vocal delivery quality in earnings conference calls. *Journal of Accounting and Economics, 80*(1), 101763. https://doi.org/10.1016/j.jacceco.2024.101763

[3]. Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research, 6*(2), 159–178. https://doi.org/10.2307/2490232

[4]. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[5]. Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium. *Journal of Accounting Research, 27(Supplement)*, 1–36. https://doi.org/10.2307/2491062

[6]. Bernard, V. L., & Thomas, J. K. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics, 13*(4), 305–340. https://doi.org/10.1016/0165-4101(90)90008-R

[7]. Chen, X., Yu, X., Chang, L., Jing, T., He, J., Wang, Z., Luo, Y., Chen, X., Liang, J., Wang, Y., & Xie, J. (2025). The sound of risk: A multimodal physics-informed acoustic model for forecasting market volatility and enhancing market interpretability. *arXiv*. https://doi.org/10.48550/arxiv.2508.18653

[8]. Da, Y., Bossa, M. N., Díaz Berenguer, A., & Sahli, H. (2024). Reducing bias in sentiment analysis models through causal mediation analysis and targeted counterfactual training. *IEEE Access*, 12, 10120–10134. https://doi.org/10.1109/access.2024.3353056

[9]. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. (2019). *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 (NAACL-HLT 2019), Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.

[10]. Du, K.; Xing, F.; Mao, R.; Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys, 56*(9), Article 220, 1–42. https://doi.org/10.1145/3649451

[11]. Ergun, Z. E.; Sefer, E. (2025). FinSentiment: Predicting financial sentiment through transfer learning*Intelligent Systems in Accounting, Finance & Management, 32*(3). https://doi.org/10.1002/isaf.70015Ewertz, J., Knickrehm, C., Nienhaus, M., & Reichmann, D. (2025). Listen closely: Measuring vocal tone in corporate disclosures. Journal of Accounting Research. https://doi.org/10.1111/1475-679X.70015

[12]. Ewertz, J.; Knickrehm, C.; Nienhaus, M.; Reichmann, D. (2026). Listen closely: Measuring vocal tone in corporate disclosures. *Journal of Accounting Research, 64*(1), 229–277. https://doi.org/10.1111/1475-679X.70015

[13]. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417. https://doi.org/10.2307/2325486

[14]. Gill, S. H., Mahar, J. A., Mahar, S. A., Razzaq, M. A., Mehmood, A., Choi, G. S., & Ashraf, I. (2026). Prosodic information extraction and classification based on MFCC features and machine learning models. *Measurement and Control, 59*(1). https://doi.org/10.1177/00202940251315031

[15]. Gupta, I. (2025). Acoustic features of corporate conference calls and market reactions (2010–2025). *SSRN*. https://doi.org/10.2139/ssrn.5607250

[16]. Hoekstra, J., & Güler, D. (2022). The mediating effect of trading volume on the relationship between investor sentiment and the return of tech companies. *Journal of Behavioral Finance*, 25, 356–373. https://doi.org/10.1080/15427560.2022.2138394

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1387

[17]. Huang, L. (2023). The impact of China economic policy uncertainty on CSI 300: An analysis of the mediating effect of investor sentiment. Advances in Economics, *Management and Political Sciences*, 51, 20230608. https://doi.org/10.54254/2754-1169/51/20230608

[18]. Huang, Y., Zhang, J., & Liu, S. (2021). Vocal tone and investor reactions: Evidence from matched earnings calls. *Review of Accounting Studies, 26*(4), 1456–1492. https://doi.org/10.1007/s11142-021-09640-7

[19]. Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research, 50*(2), 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

[20]. Li, S.; Tang, H. (2024). Multimodal alignment and fusion: A survey. arXiv. https://doi.org/10.48550/arXiv.2411.17040

[21]. Livnat, J., & Mendenhall, R. R. (2006). Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. Journal of *Accounting Research, 44*(1), 177–205. https://doi.org/10.1111/j.1475-679X.2006.00196.x

[22]. Lundberg, S. M., & Lee, S.-I. (2017). A *unified approach to interpreting model predictions*. In Advances in Neural Information Processing Systems (NeurIPS 2017, Vol. 30). Curran Associates. https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[23]. Mai, Z.; Zhang, J.; Xu, Z.; Xiao, Z. (2024). *Financial sentiment analysis meets LLaMA 3: A comprehensive analysis*. In Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI '24) (pp. 171–175). Association for Computing Machinery. https://doi.org/10.1145/3696271.3696299

[24]. Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review, 86*(4), 1383–1414. https://doi.org/10.2308/accr-10034

[25]. Mayew, W. J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *The Journal of Finance, 67*(1), 1–43. https://doi.org/10.1111/j.1540-6261.2011.01705.x

[26]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144)*. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[27]. Ruan, L., & Jiang, H. (2025). Stock price prediction using FinBERT-enhanced sentiment with SHAP explainability and differential privacy. *Mathematics, 13*(17), 2747. https://doi.org/10.3390/math13172747

[28]. Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

[29]. Todd, A.; Bowden, J.; Moshfeghi, Y. (2024). Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. *Intelligent Systems in Accounting, Finance & Management, 31*(1), e1549. https://doi.org/10.1002/isaf.1549

[30]. Wang, Z., Li, Y., & Zhang, H. (2023). *Counterfactual multimodal modeling in financial communication analysis*. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 890–902. https://doi.org/10.18653/v1/2023.emnlp-main.78

[31]. Yamaguchi, R., & Yanai, K. (2025). *Exploring cross-attention maps in multi-modal diffusion transformers for training-free semantic segmentation*. In M. Cho, I. Laptev, D. Tran, A. Yao, & H. B. Zha (Eds.), Computer vision – ACCV 2024 workshops (Lecture Notes in Computer Science, Vol. 15482). Springer. https://doi.org/10.1007/978-981-96-2641-0_18

[32]. Zhang, L., Wu, Z., & Jin, Y. (2022). Structural causal modeling for vocal sentiment in earnings calls. *Journal of Financial Markets*, 58, 100745. https://doi.org/10.1016/j.finmar.2021.100745