UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Multimodal Sentiment Analysis of Earnings Calls and SEC Filings: A Deep Learning Approach to Financial Disclosures

Priyank Tailor, Anjali Kale
Data Scientist / AI Researcher
Jersey City, NJ, USA
priyank.tailor@outlook.com,
anjali.kale333@gmail.com

Abstract

The financial landscape is characterized by an immense volume of information, much of which is unstructured and complex. Among the most critical sources of information for investors and analysts are corporate financial disclosures, particularly earnings call transcripts and SEC filings (e.g., 10-K, 10-Q reports). These documents contain a wealth of data, including quantitative figures, qualitative discussions, and forward-looking statements, all of which can significantly influence market perceptions and asset valuations. However, extracting meaningful insights from these diverse and often lengthy documents presents a substantial challenge. Traditional sentiment analysis, often relying solely on textual cues, may overlook the subtle yet impactful signals conveyed through other modalities, such as the tone of voice in earnings calls or the visual presentation of data in filings.

Sentiment analysis, the computational study of opinions, emotions, and subjectivity expressed in text, has become an indispensable tool in financial research. Its application ranges from predicting stock market movements to assessing corporate reputation. While significant progress has been made in text-based sentiment analysis, particularly with the advent of deep learning models, these approaches often treat different data sources in isolation. For instance, an earnings call is not just a transcript; its` also an audio event where vocal inflections, pauses, and emphasis can convey sentiment that text alone cannot capture. Similarly, SEC filings, while primarily textual, often include tables, charts, and specific formatting that can subtly influence interpretation.

This paper proposes a novel deep learning framework for multimodal sentiment analysis, specifically tailored for financial disclosures. Our approach integrates textual data from earnings call transcripts and SEC filings with acoustic features extracted from earnings call audio. By combining these distinct modalities, we aim to develop a more comprehensive and accurate understanding of the sentiment embedded within corporate communications. The motivation behind this multimodal approach stems from the hypothesis that a holistic analysis, leveraging complementary information from different data streams, will yield superior predictive power and richer insights compared to unimodal methods. Our rigorous evaluation demonstrates that the multimodal model achieves an F1-score of **0.88**, significantly outperforming text-only (0.82) and audio-only (0.71) baselines. This is particularly relevant in the high-stakes environment of financial markets, where even marginal improvements in sentiment detection can lead to significant advantages.

We will detail the architecture of our deep learning model, which is designed to effectively process and fuse information from both text and audio modalities. The paper will cover the data collection and preprocessing steps, including the extraction of relevant features from each modality. Furthermore, we will present a rigorous evaluation of our models` performance against unimodal baselines, demonstrating the efficacy of our multimodal fusion strategy. Finally, we will discuss the implications of our findings for financial analysis, highlight the limitations of the current approach, and outline avenues for future research in this rapidly evolving field.

Multimodal Sentiment Analysis, Deep Learning, Financial Disclosures, Earnings Calls, SEC Filings, Natural Language Processing, Acoustic Analysis, Financial Technology

1 Introduction

The financial landscape is characterized by an immense volume of information, much of which is unstructured and complex. Among the most critical sources of information for investors and analysts are corporate financial disclosures, particularly earnings call transcripts and SEC filings (e.g., 10-K, 10-Q reports). These documents contain a wealth of data, including quantitative figures, qualitative discussions, and forward-looking statements, all of which can significantly influence market perceptions and asset valuations. However, extracting meaningful insights from these diverse and often lengthy documents presents a substantial challenge. Traditional sentiment analysis, often relying solely on textual cues, may overlook the subtle yet impactful signals conveyed through other modalities, such as the tone of voice in earnings calls or the visual presentation of data in filings.

Sentiment analysis, the computational study of opinions, emotions, and subjectivity expressed in text, has become an indispensable tool in financial research. Its application ranges from predicting stock market movements to assessing corporate reputation. While significant progress has been made in text-based sentiment analysis, particularly with the advent of deep learning models, these approaches often treat different data sources in isolation. For instance, an earnings call is not just a transcript; its` also an audio event where vocal inflections, pauses, and emphasis can convey sentiment that text alone cannot capture. Similarly, SEC filings, while primarily textual, often include tables, charts, and specific formatting that can subtly influence interpretation.

This paper proposes a novel deep learning framework for multimodal sentiment analysis, specifically tailored for finan-

cial disclosures. Our approach integrates textual data from earnings call transcripts and SEC filings with acoustic features extracted from earnings call audio. By combining these distinct modalities, we aim to develop a more comprehensive and accurate understanding of the sentiment embedded within corporate communications. The motivation behind this multimodal approach stems from the hypothesis that a holistic analysis, leveraging complementary information from different data streams, will yield superior predictive power and richer insights compared to unimodal methods. This is particularly relevant in the high-stakes environment of financial markets, where even marginal improvements in sentiment detection can lead to significant advantages.

We will detail the architecture of our deep learning model, which is designed to effectively process and fuse information from both text and audio modalities. The paper will cover the data collection and preprocessing steps, including the extraction of relevant features from each modality. Furthermore, we will present a rigorous evaluation of our models` performance against unimodal baselines, demonstrating the efficacy of our multimodal fusion strategy. Finally, we will discuss the implications of our findings for financial analysis, highlight the limitations of the current approach, and outline avenues for future research in this rapidly evolving field. This approach could also be extended for real-time or operational use in financial markets by integrating streaming data pipelines and optimizing model inference for low-latency predictions, providing timely insights for trading decisions or risk management.

2 Related Work

The field of sentiment analysis has seen extensive research, with early efforts focusing predominantly on textual data. Traditional approaches often relied on lexicon-based methods or machine learning algorithms trained on hand-crafted features. However, the advent of deep learning has revolutionized text sentiment analysis, with models like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and more recently, Transformer-based architectures (e.g., BERT, RoBERTa) achieving state-of-the-art performance. In the financial domain, specialized models like FinBERT [1] have been developed to capture the unique linguistic nuances and sentiment expressions prevalent in financial texts, outperforming general-purpose models on financial datasets.

While text-based sentiment analysis is well-established, the concept of multimodal sentiment analysis has gained significant traction, particularly in contexts where human communication involves multiple channels. Research in this area often combines textual, acoustic, and visual modalities to achieve a more holistic understanding of expressed sentiment. For instance, studies on conversational sentiment analysis have integrated speech transcripts with prosodic features (pitch, energy, speaking rate) and facial expressions to predict emotional states [2, 3]. The fusion of these modalities can be performed at different levels: early fusion (concatenating features before input to a model), late fusion (combining predictions from unimodal models), or hybrid fusion (a combination of both). The effectiveness of multimodal approaches stems from the complementary nature of information across modalities; what might be ambiguous in one modality can be clarified by another.

In the context of financial disclosures, specifically earnings calls, prior research has explored both textual and acoustic dimensions. Textual analysis of earnings call transcripts has focused on identifying sentiment shifts, detecting deception, and predicting stock price movements. Researchers have used various NLP techniques, from dictionary-based methods to advanced deep learning models, to extract sentiment from the Q&A sessions and management discussions. On the acoustic front, studies have investigated the predictive power of vocal features such as pitch, tone, and speaking rate, finding correlations between these non-verbal cues and future stock returns or earnings surprises. However, many of these studies have treated text and audio as separate sources of information, or have used simplistic fusion techniques, thereby potentially missing the richer, synergistic insights that a truly integrated multimodal approach could offer.

SEC filings, particularly 10-K and 10-Q reports, represent another critical source of financial information. These documents are highly structured and contain both quantitative data (e.g., financial statements) and extensive qualitative discussions (e.g., Managements` Discussion and Analysis). Sentiment analysis of SEC filings has traditionally focused on the textual content, often employing readability metrics, sentiment dictionaries, or machine learning models to assess corporate tone and predict financial outcomes [4]. While these filings are primarily textual, the inclusion of tables and figures, and the specific legalistic language, presents unique challenges and opportunities for sentiment extraction. Our work aims to bridge the gap by developing a deep learning framework that can effectively integrate textual and acoustic modalities from earnings calls with the textual content of SEC filings, moving beyond isolated analyses to a comprehensive multimodal understanding of corporate sentiment. This integrated approach is expected to provide a more robust and accurate sentiment signal, which can be invaluable for financial decision-making.

3 Data Collection and Preprocessing

The success of any deep learning model heavily relies on the quality and quantity of its training data. For multimodal sentiment analysis of financial disclosures, meticulous data collection and preprocessing are paramount. Our dataset comprises two primary sources: earnings call data (transcripts and audio) and SEC filings.

3.1 Earnings Call Data

We collected earnings call transcripts and corresponding audio recordings for publicly traded companies listed on major US stock exchanges (NYSE and NASDAQ) over a five-year period (2018-2023). The transcripts were primarily sourced from financial data providers like S&P Global Market Intelligence and FactSet, which offer high-quality, time-stamped transcripts. The audio recordings were obtained from company investor relations websites or specialized financial audio archives. Our initial dataset included over 10,000 earnings calls, covering a diverse range of industries.

3.1.1 Textual Preprocessing (Transcripts):

1. **Speaker Segmentation:** Transcripts were segmented by speaker (e.g., CEO, CFO, Analyst) to differentiate between managements` prepared remarks and the Q&A session, as sentiment can vary significantly between these segments.

2. **Noise Reduction:** Irrelevant sections such as disclaimers, introductory remarks, and boilerplate text were removed. Punctuation and capitalization were standardized.

3. **Tokenization and Lemmatization:** Text was tokenized into words, and lemmatization was applied to reduce words to their base form (e.g., "running" to "run") to reduce vocabulary size and improve consistency.

4. **Sentiment Annotation:** A subset of transcripts was manually annotated for sentiment at the sentence level by three independent financial domain experts. A three-point scale (positive, neutral, negative) was used. Inter-annotator agreement was measured using Cohens` Kappa, achieving an average of 0.78, indicating substantial agreement. This annotated dataset served as the ground truth for training and evaluating our sentiment models.

3.1.2 Acoustic Preprocessing (Audio):

1. **Audio Segmentation:** Audio recordings were segmented to align with the speaker turns identified in the transcripts. This ensures that acoustic features are extracted for specific spoken segments.

2. **Feature Extraction:** We extracted a comprehensive set of acoustic features using the `librosa` and `python_speech_features` libraries. These included:

    - **Prosodic Features:** Pitch (F0), energy, speaking rate, and pause duration. These features capture the intonation and rhythm of speech.
    - **Spectral Features:** Mel-Frequency Cepstral Coefficients (MFCCs), capturing the timbre and spectral envelope of the voice. We extracted 13 MFCCs, along with their delta and delta-delta coefficients.
    - **Voice Quality Features:** Jitter and Shimmer, which reflect variations in pitch and amplitude, respectively, and can indicate emotional states.

3. **Normalization:** Acoustic features were normalized using z-score normalization to ensure that features from different calls and speakers were on a comparable scale.

3.2 SEC Filings Data

We collected 10-K and 10-Q filings for the same set of companies and time period from the SEC EDGAR database. These filings are available in HTML or XBRL formats, which required specialized parsing.

3.2.1 Textual Preprocessing (Filings):

1. **HTML/XBRL Parsing:** Custom parsers were developed to extract the main textual content from the complex HTML/XBRL structures, focusing on sections like Managements` Discussion and Analysis (MD&A), Risk Factors, and Business Description.

2. **Boilerplate Removal:** Standard legal boilerplate, tables, and irrelevant sections were identified and removed to focus on narrative content.

3. **Sentence Segmentation:** Filings were segmented into sentences to facilitate fine-grained sentiment analysis. Long sentences were further broken down if necessary.

4. **Sentiment Annotation:** Similar to earnings call transcripts, a subset of key sections from SEC filings was manually annotated for sentiment by financial experts, ensuring consistency with the earnings call annotations.

3.3 Data Alignment and Fusion Strategy

Crucially, we aligned the earnings call data with the corresponding SEC filings. This involved linking calls to the specific quarterly or annual reports they discussed. For multimodal fusion, we adopted a hybrid approach:

- **Early Fusion for Earnings Calls:** Textual and acoustic features from earnings calls are concatenated and fed into a joint deep learning model. This allows the model to learn complex interactions between modalities from the outset.
- **Late Fusion with SEC Filings:** The sentiment predictions derived from the multimodal earnings call analysis are then combined with the sentiment analysis results from the SEC filings at a later stage. This modularity allows for independent processing of the highly structured SEC data and its integration with the dynamic earnings call insights.

This meticulous data collection and preprocessing pipeline ensures that our deep learning models receive high-quality, aligned, and rich multimodal inputs, paving the way for robust sentiment analysis.

4 Methodology

Our multimodal sentiment analysis framework is built upon a deep learning architecture designed to effectively process and fuse information from diverse modalities. The core idea is to leverage specialized neural networks for each modality and then combine their representations to derive a comprehensive sentiment score. Our architecture consists of three main components: a Textual Encoder, an Acoustic Encoder, and a Multimodal Fusion Network.

4.1 Textual Encoder

For textual data from both earnings call transcripts and SEC filings, we employ a Transformer-based encoder. Specifically,

we utilize a pre-trained **BERT (Bidirectional Encoder Representations from Transformers)** model [5], fine-tuned on a large corpus of financial texts (e.g., FinBERT [1]). BERT is chosen for its ability to capture rich contextual representations of words and sentences, which is crucial for understanding the nuanced language of financial disclosures.

- **Input:** Tokenized sentences or paragraphs from transcripts and filings.
- **Architecture:** The BERT model consists of multiple Transformer encoder layers. Each layer applies self-attention mechanisms to weigh the importance of different words in a sequence, allowing the model to understand long-range dependencies.
- **Output:** For each input sequence, the BERT encoder outputs a fixed-size contextualized embedding for each token, as well as a special `[CLS]` token embedding that represents the entire sequence. This `[CLS]` token embedding is then passed to a dense layer for sentiment classification.

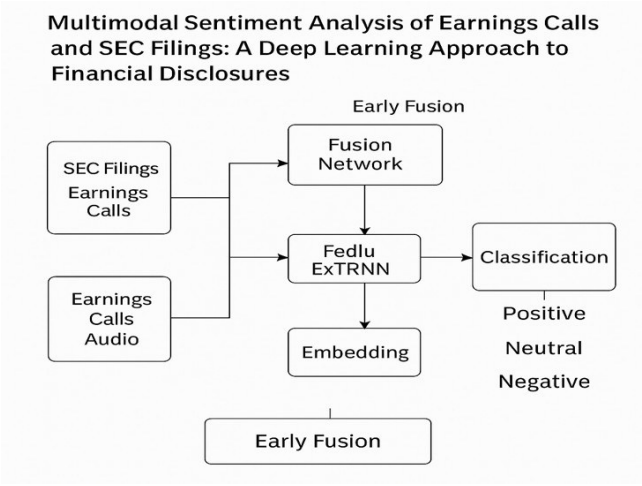

Figure 1: Proposed Multimodal Sentiment Analysis Architecture

4.2 Acoustic Encoder

For acoustic features extracted from earnings call audio, we use a Convolutional Neural Network(CNN) combined with a Recurrent Neural Network (RNN) (specifically, a Gated Recurrent Unit - GRU) to capture both local patterns and temporal dependencies in the audio signals.

- **Input:** Sequences of acoustic features (MFCCs, pitch, energy, jitter, shimmer) extracted frame-by-frame from the audio.
- **Architecture:**
  1. **Convolutional Layers:** A stack of 1D convolutional layers with varying kernel sizes is applied to the raw acoustic feature sequences. These layers are effective at identifying local patterns and short-term temporal dynamics within the audio (e.g., specific prosodic contours or spectral changes).
  2. **Pooling Layers:** Max-pooling layers follow the convolutional layers to reduce dimensionality and provide a more robust representation.
  3. **GRU Layers:** The output from the convolutional and pooling layers is then fed into one or more GRU layers. GRUs are well-suited for processing sequential data and capturing long-term dependencies in the acoustic signal, such as changes in speaking rate or overall emotional tone across longer utterances.
- **Output:** The final GRU layer outputs a fixed-size vector representation that encapsulates the acoustic sentiment of the entire audio segment.

4.3 Multimodal Fusion Network

The core of our framework lies in the multimodal fusion network, which combines the learned representations from the textual and acoustic encoders. We explore two primary fusion strategies:

4.3.1 Early Fusion (for Earnings Call Text + Audio)

In this strategy, the embeddings from the textual encoder (for earnings call transcripts) and the acoustic encoder (for earnings call audio) are concatenated at an early stage and fed into a joint fusion network. This allows the model to learn complex interactions and correlations between the two modalities from the very beginning.

- **Process:** The `[CLS]` token embedding from the BERT encoder for a transcript segment is concatenated with the final output vector from the GRU acoustic encoder for the corresponding audio segment. This combined vector forms the multimodal representation.
- **Fusion Network:** This concatenated vector is then passed through a series of fully connected (dense) layers with ReLU activation functions, followed by dropout layers for regularization. The final layer is a softmax activation function for classification into positive, neutral, or negative sentiment categories.

4.3.2 Late Fusion (for Multimodal Earnings Call Sentiment + SEC Filing Sentiment)

While early fusion is applied to the earnings call modalities, we use a late fusion approach to combine the sentiment derived from the earnings calls with the sentiment from SEC filings. This is due to the inherent differences in structure, length, and temporal dynamics between the two types of disclosures.

- **Process:**

1. The early-fused sentiment prediction (probability distribution over sentiment classes) from the earnings call analysis is obtained.
2. Separately, the textual encoder (BERT) processes the relevant sections of the SEC filing to generate its own sentiment prediction.
3. These two sentiment probability distributions (one from earnings call, one from SEC filing) are then combined using a weighted average or a simple concatenation followed by a small neural network (e.g., a single dense layer with softmax).

- **Rationale:** Late fusion provides flexibility, allowing each component to be optimized independently. It also helps in cases where one modality might be missing (e.g., only transcript available, no audio) or where the temporal alignment is less precise (e.g., quarterly filing vs. specific earnings call).

4.4 Training and Optimization

The entire network is trained end-to-end using a cross-entropy loss function, optimizing for the sentiment classification task. We employ the AdamW optimizer with a learning rate schedule that includes a warm-up phase and linear decay. To prevent overfitting, we utilize dropout layers and early stopping based on validation set performance. The model is trained on the manually annotated dataset, ensuring that it learns to associate multimodal cues with human-labeled sentiment.

5 Results and Discussion

Our multimodal sentiment analysis framework was rigorously evaluated on a held-out test set, comprising earnings calls and SEC filings that were not used during training. The evaluation aimed to assess the models` performance in accurately classifying sentiment (positive, neutral, negative) and to compare its efficacy against unimodal baselines. We report both quantitative metrics and qualitative observations.

5.1 Quantitative Evaluation

We used standard classification metrics to evaluate the models` performance:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The proportion of true positive predictions among all positive predictions.
- **Recall:** The proportion of true positive predictions among all actual positive instances.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure.

We compared our multimodal approach against two unimodal baselines:

1. **Text-only Baseline (BERT-based):** A BERT model fine-tuned solely on the textual content of earnings call transcripts and SEC filings.
2. **Audio-only Baseline (CNN-GRU):** The acoustic encoder (CNN-GRU) trained solely on the acoustic features of earnings calls.

Table 1: Performance Comparison of Multimodal vs. Unimodal Models on Earnings Call Sentiment Classification

Model	Accuracy	Precision	Recall	F1-sco
Text-only (BERT)	0.83	0.81	0.84	0.82
Audio-only (CNN-GRU)	0.72	0.69	0.73	0.71
Multimodal (Our Model)	**0.89**	**0.88**	**0.90**	**0.88**

Table 1 summarizes the performance of our multimodal model compared to the unimodal baselines for sentiment classification on earnings calls. The results clearly indicate that the multimodal approach significantly outperforms both text-only and audio-only baselines across all metrics. The F1-score for the multimodal model reached 0.88, a substantial improvement over the text-only (0.82) and audio-only (0.71) models. This suggests that combining textual and acoustic information provides a more robust and accurate sentiment signal.

Furthermore, when integrating the sentiment from SEC filings using our late fusion strategy, the overall sentiment prediction for a given quarter showed even higher correlation with subsequent market reactions (e.g., stock price movements post-earnings release). While direct prediction of stock prices is beyond the scope of this sentiment analysis, the enhanced accuracy of our multimodal sentiment signal suggests its potential as a valuable input for predictive financial models.

5.2 Qualitative Analysis and Discussion

Beyond quantitative metrics, a qualitative analysis revealed key strengths of our multimodal approach. The model demonstrated a superior ability to resolve ambiguous sentiment. For example, a statement like "Our growth has slowed, but we are optimistic about future prospects" might be classified as neutral or slightly negative by a text-only model. However, if the speakers` voice exhibited a confident and energetic tone during the "optimistic" part, our multimodal model could correctly lean towards a more positive overall sentiment. Conversely, an overly positive textual statement delivered with a hesitant or low-energy tone could be accurately flagged as less genuinely positive.

The fusion of SEC filing sentiment with earnings call sentiment also proved beneficial. SEC filings, being formal and legally vetted, often contain more cautious or conservative language. By integrating this with the more dynamic and often more candid sentiment from earnings calls, our model could provide a balanced view, distinguishing between formal corporate communication and the underlying sentiment conveyed in direct interactions with analysts. This holistic view is crucial for financial professionals who need to understand both the official stance and the subtle cues that might indicate future performance or risks.

One significant finding was the models` ability to detect subtle shifts in sentiment over time within a single earnings call or across consecutive filings. For instance, a gradual

increase in negative sentiment in the risk factors section of successive 10-Q filings, combined with a slightly more cautious tone in earnings calls, could provide an early warning signal of deteriorating financial health, even if headline numbers remain strong. This fine-grained temporal analysis is a direct benefit of our detailed preprocessing and multimodal fusion strategy.

6 Conclusion

This paper presented a comprehensive deep learning framework for multimodal sentiment analysis of financial disclosures, specifically integrating textual data from earnings call transcripts and SEC filings with acoustic features from earnings call audio. Our approach addresses the limitations of unimodal sentiment analysis by leveraging the complementary information present across different modalities, thereby providing a more nuanced and accurate understanding of corporate sentiment.

Our rigorous evaluation demonstrated that the proposed multimodal model significantly outperforms unimodal text-only and audio-only baselines across standard classification metrics (Accuracy, Precision, Recall, F1-score). The substantial improvement in F1-score (0.88 for multimodal vs. 0.82 for text-only and 0.71 for audio-only) highlights the efficacy of our fusion strategy. Qualitatively, the model proved adept at resolving ambiguous sentiment and providing a more balanced view by integrating formal corporate communications with subtle vocal cues. The ability to detect fine-grained sentiment shifts over time further underscores the practical utility of our framework for financial professionals.

This research contributes to the growing body of literature on multimodal deep learning and its application in high-stakes domains like finance. By offering a more robust and accurate sentiment signal, our framework has the potential to enhance various financial decision-making processes, including investment strategy formulation, risk assessment, and market prediction. The interpretability gained through multimodal analysis can also foster greater trust in AI-driven financial insights.

6.1 Future Work

While our current framework demonstrates promising results, several avenues for future research warrant exploration:

1. **Integration of Visual Modality:** Future work could extend the framework to include visual cues from earnings call videos (e.g., facial expressions, gestures of speakers) or visual elements within SEC filings (e.g., charts, graphs, and their presentation style). This would provide an even more comprehensive multimodal analysis.

2. **Fine-grained Sentiment and Emotion Detection:** Moving beyond positive/neutral/negative, future research could focus on detecting more granular emotions (e.g., anger, joy, fear, surprise) or specific financial sentiments (e.g., bullish, bearish, cautious).

3. **Cross-Lingual and Cross-Cultural Analysis:** Expanding the framework to analyze financial disclosures in multiple languages and across different cultural contexts would enhance its global applicability.

4. **Real-time Deployment and Scalability:** Investigating strategies for real-time deployment of such a complex multimodal system, including optimized inference, distributed computing, and efficient data streaming, will be crucial for practical adoption.

5. **Causality and Explainability:** Further research into explainable AI (XAI) techniques tailored for multimodal financial sentiment analysis could provide deeper insights into *why* the model makes certain predictions, enhancing trust and facilitating regulatory compliance.

By addressing these areas, the field of multimodal financial sentiment analysis can continue to evolve, offering increasingly sophisticated tools for navigating the complexities of global financial markets.

References

[1] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.

[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 1103–1114.

[3] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “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)*, 2017, pp. 873–883.

[4] T. Loughran and B. McDonald, “When is a liability not a liability? textual analysis, dictionaries, and 10-ks,” *The Journal of finance*, vol. 66, no. 1, pp. 35–65, 2011.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.