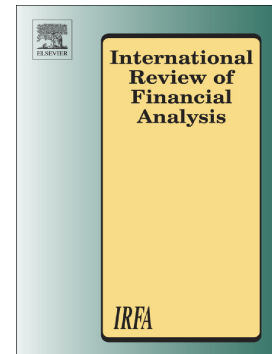


# Journal Pre-proof

A multimodal sentiment classifier for financial decision making

Andrew Todd, James Bowden, Mark Cummins, Yang Su



PII: S1057-5219(25)00409-0

DOI: <https://doi.org/10.1016/j.irfa.2025.104322>

Reference: FINANA 104322

To appear in: *International Review of Financial Analysis*

Received date: 20 November 2024

Revised date: 27 April 2025

Accepted date: 6 May 2025

Please cite this article as: A. Todd, J. Bowden, M. Cummins, et al., A multimodal sentiment classifier for financial decision making, *International Review of Financial Analysis* (2024), <https://doi.org/10.1016/j.irfa.2025.104322>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier Inc.

# A Multimodal Sentiment Classifier for Financial Decision Making

Andrew Todd<sup>a,~</sup>, James Bowden<sup>a</sup>, Mark Cummins<sup>a</sup>, Yang Su<sup>b</sup>

<sup>a</sup> Strathclyde Business School, University of Strathclyde, 199 Cathedral St, Glasgow G4 0QU, UK

<sup>b</sup> Trinity Business School, Trinity College Dublin, 182 Pearse St, Dublin 2, D02 F6N2, Ireland

**Abstract:** This study pioneers a multimodal approach to financial sentiment analysis through the integration of audio and textual data to enhance predictive accuracy. Motivated by the underutilisation of paralinguistic features and deep learning techniques in financial sentiment analysis, we introduce a novel deep learning-enabled multimodal classifier trained on corporate earnings calls using a subset of S&P 100 constituents. Our methodology incorporates FinBERT, a financial variant of Bidirectional Encoder Representation Transformations (BERT), alongside paralinguistic features and a deep learning classifier. Comparative analysis against established sentiment analysis methods, including dictionary approaches and machine learning models, suggests that our multimodal classifier achieves improved out-of-sample accuracy. Specifically, the inclusion of paralinguistic characteristics improves sentiment detection accuracy. Our research provides nuanced insights into sentiment analysis detection of different speakers (managers and analysts) during both the management discussion and Q&A sections of corporate earnings calls. Combined, our results suggest that multimodal sentiment analysis classification possesses the ability to deepen our understanding of the interplay between sentiment and market characteristics.

**Keywords:** sentiment analysis, multimodal analysis, transformer model, deep learning, earnings call

\*This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

~Corresponding author.

Email addresses (ORCID ID): andrew.todd@strath.ac.uk (0000-0001-7440-2342),

james.bowden@strath.ac.uk, mark.cummins@strath.ac.uk, sua@tcd.ie

Conflict of Interests: Andrew Todd, James Bowden, Mark Cummins, Yang Su – None.

## 1. Introduction

Over the last two decades, researchers have increasingly used various sources of textual data to determine statistically and economically significant drivers of financial market activity, such as stock returns, trading volumes, and price volatility. Sentiment analysis, in particular, has emerged as a prominent tool for extracting subjective information from unstructured financial texts (Soleymani et al., 2017). Early applications of sentiment analysis methods, for example Antweiler and Frank (2004), have inspired a wide-spanning field of analysis which incorporates various sources of qualitative

information, such as company filings, public news and social media. The literature now encompasses several sentiment analysis methods ranging from traditional dictionary approaches to more computationally demanding machine learning (ML) and deep learning (DL), with recent evidence suggesting that, in certain contexts, advanced ML and DL approaches may outperform traditional dictionary-based methods in capturing financial sentiment (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017).

Despite these advances, El-Haj et al., (2019) identify that the application of Natural Language Processing (NLP) for sentiment classification to support financial decision making is less advanced than other disciplines, with the majority of studies applying ML techniques using traditional, long-established classifiers, such as Naive Bayes (Antweiler and Frank, 2004; Li, 2010; Sprenger et al., 2013). While Naive Bayes and other similar algorithms are less computationally demanding and have been shown to classify financial sentiment with a considerable degree of accuracy, alternative approaches such as DL have been shown to potentially improve sentiment classification accuracy across several non-financial domains (Munikar, Shakya and Shrestha, 2019; Sun et al., 2020; Alamoudi and Alghamdi, 2021).

Moreover, a limitation of existing sentiment analysis methods in financial sentiment detection is the predominant focus on the textual modality, with comparatively less focus on potentially informative audio cues. A notable exception is Mayew and Venkatachalam (2012), who leverage vocal cues to assess managers' affective states and evaluate associations with securities pricing behaviour, finding vocal cues to be effective indicators of emotion. Prior experimental evidence by Mehrabian and Wiener (1967) and Mehrabian and Ferris (1967) suggests that, when verbal and non-verbal cues are inconsistent, listeners rely more heavily on facial expressions and vocal tone than the literal meaning of words. Combined, these findings highlight the potential for multimodal sentiment analysis, incorporating vocal and/or visual modalities alongside text, to more closely approximate the multidimensional nature of human communication. While the availability of financial disclosures containing all three modalities remains limited, earnings conference calls represent a useful dual modality (audio-textual) setting for financial sentiment analysis.

Our research investigates the extent to which combining paralinguistic audio features (such as pitch, intensity and audio length) with textual analysis, within a deep learning framework, improves sentiment classification accuracy in earnings calls, and whether such improvements enhance the measurement of sentiment relevant to financial decision-making contexts. To do so, we establish a DL-enabled multimodal sentiment classifier which leverages a financial version of Bidirectional Encoder

Representation Transformations (BERT)<sup>1</sup>, *FinBERT*, alongside paralinguistic features and a deep neural network (DNN) classifier, and which is trained on a sample of corporate earnings calls concerning S&P 100 firms. We then benchmark the performance of our multimodal DNN classifier against traditional dictionary methods, Naïve Bayes classifiers, and text-only transformer-based models such as BERT and FinBERT. In this respect, we build on the disclosure sentiment comparison of dictionary and machine learning methods established by Frankel et al., (2022), Renault (2020), and Kearney and Liu (2014).

We find that the integration of audio features can enhance sentiment classification accuracy, even relative to advanced text-only deep learning models, within the context of corporate earnings calls. This improved measurement of corporate disclosure tone may allow for a deeper understanding of the dynamics between sentiment and market reactions and investment decisions, reflecting a promising direction for research in financial sentiment analysis (Todd et al., 2023). While prior research from the computer science domain has combined text and audio modalities to improve sentiment classification (Houjeij et al., 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021), comparatively few studies have explored such methods within a finance context. Indeed, to the best of our knowledge, this study is amongst the first to apply a DL-based multimodal classifier to financial disclosure data for the purposes of sentiment classification.

Our analysis highlights specific components of earnings conference calls, such as the management discussion and question-and-answer (Q&A) sections, where sentiment classification varies substantially. Specifically, we find that multimodal models incorporating text and audio features outperform traditional text-only approaches, including dictionary-based methods and Naïve Bayes classifiers. On our validation dataset, the multimodal FinBERT DNN achieves out-of-sample accuracy of 74.88%, while traditional classifiers achieve accuracy rates between 42.18% and 56.40%. Moreover, our multimodal model also improves upon recent transformer-based models, outperforming both BERT (65.17%) and FinBERT (73.46%). These results suggest that incorporating audio features provides incremental performance gains even relative to advanced text-based deep learning models.

When we isolate the audio modality, we find the accuracy rate to be comparatively poor (34.12%), suggesting that accuracy is primarily determined by the textual modality, but is enhanced by the inclusion of audio characteristics. Our multimodal classifier performs comparatively well in identifying positive and negative sentiments, with neutral sentiment classification comparable to that of recent transformer-based models. Performance is notably stronger when classifying manager sentiment during the Q&A section of the call, a segment that is rich in affective cues (McKay Price et al., 2010).

---

<sup>1</sup> See Devlin et al., (2019) for an in-depth explanation of this transformer model and Yang et al., (2020) for an explanation of FinBERT - the financial version of BERT.

However, our classifier is dominated by BERT and FinBERT at classifying sentiment during the management discussion aspect of the call. Traditional classifiers exhibit lower accuracy across sentiment categories in our sample, consistent with prior evidence on their limitations in nuanced financial text (Kearney and Liu, 2014).

Combined, our findings indicate that the incorporation of audio-level data into deep learning-based sentiment analysis can improve classifier accuracy relative to existing text-only models, at least within the context of corporate earnings calls, and thus enhance the robustness of sentiment measures used in future finance research. This may allow for more accurate modelling of the association between disclosure tone and financial market activity. This study also demonstrates the viability of applying multimodal techniques in finance, therefore providing a foundation for future research to apply similar multimodal frameworks to other financial settings beyond corporate earnings calls. The remainder of the paper is organised as follows. Section 2 identifies the stages involved in the development of our aligned audio-textual dataset and provides an overview of the data used for this comparative analysis. Section 3 describes our multimodal classifier design, and outlines each of the benchmark sentiment analysis methods used for comparison. Section 4 discusses the accuracy results of our multimodal classifier, with additional testing, before Section 5 concludes.

## 2. Data

For our analysis, we take a snapshot of S&P 100 constituents in 2021 and use the financial database *Finnhub* to obtain corporate earnings call transcripts and accompanying audio recordings for each firm. Company earnings calls were then downloaded for the sample period beginning in 2005 and ending in 2021. Due to the manual verification required to ensure high-quality alignment between textual and audio data at the sentence level, we restrict the sample size used for our model training to twenty constituents of the S&P100 index, selecting four firms at random from each quintile when ranking all companies by market capitalisation. Through this random selection, we ensure that model training is conducted across the firm size distribution and represents a diverse range of firms within each size category, supporting the generalisability of the model. Thus, our final dataset consists of text and audio files corresponding to corporate earnings call transcripts for twenty sample firms across the seventeen-year sample period. While not all firms were continuous members of the S&P100 index throughout the sample period, our sampling approach ensures consistency based on 2021 membership. We acknowledge that this approach may not fully eliminate selection bias. However, it supports our objective of developing a multimodal sentiment classifier trained on a contemporary and representative cross-section of major U.S. firms' financial disclosures.

An overview of the twenty firms, including the operating industry and proportionate weighting within the sample, is provided in Appendix 1, and a comparison of the industry weight of our sample

versus the broader S&P100 index can be found in Appendix 2. Broadly, the industry weightings of our sample are consistent with the S&P100 index, with some exceptions: we are comparatively overweight for the financial (+7%) and healthcare (+5%) sectors, while underweight the consumer discretionary sector (-18%). It is important to note that the choice of sample may influence both the linguistic and paralinguistic features in the training data, given that firms operating in more heavily regulated sectors, such as financials and healthcare may adopt more formal and technical language (Loughran and McDonald, 2011). Though our classifier is broadly similar, in terms of characteristics, with the S&P100, some caution may be warranted when generalising performance within those industry sectors that are less represented by our dataset.

Next, we focus on the generation of paralinguistic features, such as tone and pitch, associated with each earnings call sentence. This process consists of two stages. Firstly, we align earnings call transcripts with the corresponding audio file to generate sentence level audio clips through a process of “iterative forced alignment” (Section 2.1). Secondly, we generate paralinguistic features from the sentence level audio clips using a widely used phonetics tool *Praat* (Section 2.2).<sup>2</sup> For the iterative forced alignment process, we follow the method previously established by Li et al. (2020), using Python to leverage the iterative forced alignment package *Aeneas*.<sup>3</sup> However, we differentiate our approach by focusing on all earnings call participants (managers and analysts) and directly aligning calls at the sentence level. Comparatively, Li et al., (2020) narrow their focus to the “Management Discussion” section of the earnings call, and adopt a two-stage alignment process, aligning first at a paragraph before subsequently aligning at the sentence level. We elaborate further on the forced alignment process below.

### 2.1. *Iterative Forced Alignment of Earnings Calls*

To enable multimodal sentiment classification at the sentence level, we align each sentence in the earnings call transcript with its corresponding segment in the audio recording. This ‘forced alignment’ process, though computationally demanding, is necessary to extract paralinguistic features (such as pitch, intensity, and audio length) on a sentence-by-sentence basis, and to pair these features directly with the corresponding sentence-level textual data. Without this intervention, the association between textual and non-textual information would be limited to broader document-level analysis which, given that there can be multiple speakers and participants on any one conference call, would likely reduce the precision of the extracted audio features, and thus our sentiment measure.

The process of iterative forced alignment involves the automatic association of an audio file with a group of transcribed text files, resulting in the estimation of timings corresponding to the presence

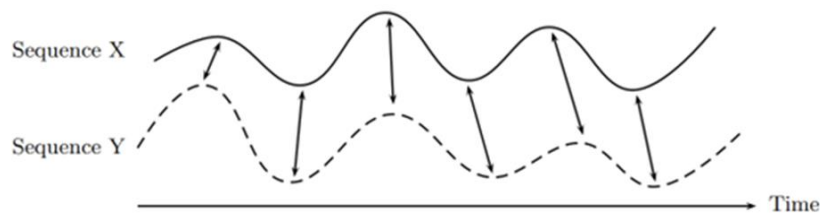
---

<sup>2</sup> Praat (<https://www.fon.hum.uva.nl/praat/>) is a free and open-source phonetics tool created by Professors Paul Boersma and David Weenink, Department of Phonetic Sciences at the University of Amsterdam.

<sup>3</sup> <https://github.com/readbeyond/aeneas>.

of each text file within the audio file. For the purposes of estimation, Aeneas adopts a Dynamic Time Warping (DTW) algorithm that identifies the optimal alignment between two time-dependent sequences (Muller, 2007), represented in Figure 1. The Aeneas package uses DTW to align the Mel-Frequency Cepstral Coefficients (MFCCs) representation of the real audio with a synthesised version created for each sentence using text-to-speech (TTS).<sup>4,5</sup> The resulting time map captures the duration of each synthesised sentence and stitches them together into a continuous synthesised wave file. MFCCs are generated by comparing audio frames to a set of cosine waveforms, yielding numerical representations of spectral similarity.<sup>6</sup> For both the real and synthesised audio files the algorithm can then map a time in the audio file corresponding to a frame in either the real or synthesised MFCC matrices.

**Figure 1.** Visualisation of the Dynamic Time Warping Process (Muller, 2007)



**Notes:** This figure gives a visual representation of DTW aligning two time-dependent series. The arrows represent the alignment points on both time-series.

The alignment process then employs the DTW algorithm to find the minimum cost path to transform the synthesised audio into the real audio.<sup>7</sup> The output from the DTW is a time map which relates the columns, or time frames, in the real MFCC matrix to the synthesised MFCC matrix. This enables the timing of each synthesised sentences to be mapped onto the real audio, identifying when each sentence in the transcript is spoken. The real audio file is then split into multiple separate audio files that relate to each sentence. Figure 2 illustrates this process, where five sentences spoken during an audio recording are converted to five smaller audio clips relating to the estimation of each sentence.

<sup>4</sup> A synthetic audio file is a file that has been created using a text-to-speech (TTS) engine. In this case it takes the sentence level text data and creates an artificial audio recording for each sentence then joins these files together to create one overall synthesised audio file.

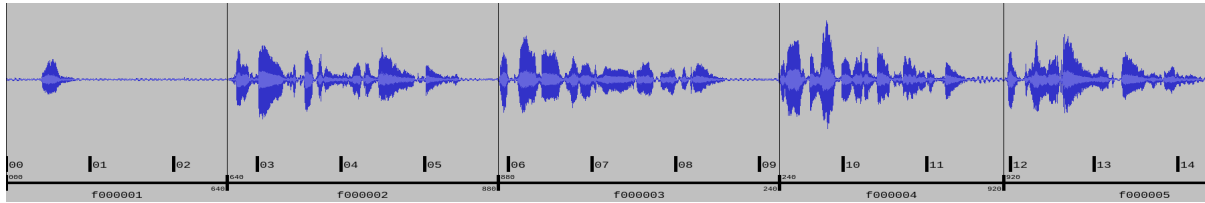
<sup>5</sup> The audio produced using TTS needs to be intelligible audio for the alignment to work but it does not necessarily need to sound natural.

<sup>6</sup> For each MFCC matrix, each column represents one frame or sub-interval of the audio file. For example, if one frame was denoted as being 1 second then the example synthesised soundwave would contain 8 frames and subsequently the matrix would contain 8 columns. As both the real wave and the synthesised wave will have different lengths, each matrix will have a different number of columns. The rows in both matrices represent one MFCC coefficient of each frame.

<sup>7</sup> To do so, Aeneas takes the dot product of both MFCC matrices to create a cost matrix and then runs the DTW algorithm over this cost matrix to find an approximation of the minimum cost path.

Finally, we use Praat’s *Parselmouth* Python library<sup>8</sup> to generate numerical representations of each sentence level clip, hence creating sentence level paralinguistic features.

**Figure 2.** An Example of Conversion from an Audio File to Sentence-Level Audio



**Notes:** This is a depiction of the final form of the real audio wave that contains time stamps relating to each sentence spoken. F000001 relates to the time in which the first sentence on the transcript is spoken and so forth.

## 2.2. Dataset Characteristics

From a total sample of 595,074 sentences, linked to the 20 firms in question, we were able to secure a usable subsample of 164,632 sentences with full audio-text alignment and available paralinguistic features. We then label a sample of 2,106 sentences, randomly drawn from this usable subsample of 164,632 sentences. This represents ~1.2% of this usable subsample. While this labelled sample is relatively small, the sentiment labelling exercise is both time consuming and resource intensive. Specifically, two of the co-authors to this study independently and separately classified each of the 2,106 sentences as conveying positive, neutral, or negative sentiment based on textual information. Only when both co-authors matched in terms of their sentiment classifications, were sentences included in the final labelled dataset. This dual-labelling approach was adopted to provide greater confidence in the reliability of the labelling and to mitigate any subjective bias in using a single-labelling approach. We ensure an equal number of positive, neutral and negative sentences (i.e. 702 sentences per category). We adopt this approach as ML and DL algorithms are known to be less predictive on unseen data when trained on imbalanced dependent variable sets, and often become biased towards the overrepresented data category (Rezapour, 2020).

Sentences were manually classified as containing “positive”, “neutral” or “negative” sentiment as follows. We classify sentences as negative (positive) if they have unfavourable (favourable) connotations towards the performance of the firm engaged in the earnings call. Sentences are classified as “neutral” if they do not contain any significant information regarding firm performance. Our results indicate that there is a slight skew towards positive messages from managers and a slight skew towards negative messages from analysts. This is consistent with prior studies finding that managers speak with significantly greater optimism than analysts on earnings call (Brockman, Li and McKay Price, 2015), perhaps due to a managers’ preference for positive framing when disseminating corporate performance

<sup>8</sup> <https://parselmouth.readthedocs.io/en/stable/>.

information. Furthermore, Renault (2017) finds that call sentiment is more positive towards the beginning of an earnings call due to managerial introductions, before becoming more balanced later in the call, when financial analysts begin questioning managers.

**Table 1.** Summary of Earnings Call Sentences

*Panel A: By Year and Call Section*

<b>Year</b>	<b>No. Sentences</b>	<b>MD (%)</b>	<b>Q&amp;A (%)</b>	<b>Avg. Sentiment</b>	<b>Avg. Audio (s)</b>
2005	8	100.00	0.00	0.50	12.75
2006	46	36.96	63.04	0.33	19.86
2007	116	41.38	58.62	-0.03	14.00
2008	176	22.73	77.27	-0.19	14.98
2009	173	8.67	91.33	-0.18	17.73
2010	135	14.81	85.19	-0.01	14.97
2011	140	1.43	98.57	-0.04	11.43
2012	152	0.66	99.34	-0.03	13.78
2013	136	1.47	98.53	0.07	15.12
2014	157	8.28	91.72	0.06	18.26
2015	153	0.00	100.00	-0.08	18.81
2016	147	1.36	98.64	-0.03	14.72
2017	114	1.75	98.25	0.17	17.58
2018	117	0.00	100.00	0.17	15.43
2019	155	0.00	100.00	0.03	19.06
2020	149	0.00	100.00	0.07	18.51
2021	32	0.00	100.00	0.13	15.72
<b>Total</b>	<b>2106</b>	<b>8.07</b>	<b>91.93</b>	<b>0.00</b>	<b>16.16</b>

**Panel B: By Call Section and Speaker**

Sentiment	N	Call Section		Speaker	
		MD (%)	Q&A (%)	Mgmt (%)	Analyst (%)
Positive	702	9.40	90.60	55.56	44.44
Negative	702	7.83	92.17	50.57	49.43
Neutral	702	6.98	93.02	54.99	45.01
<b>Total</b>	<b>2106</b>	<b>8.07</b>	<b>91.93</b>	<b>53.70</b>	<b>46.30</b>

**Notes:** This table disaggregates the extracted 2,106-sentence sample across the years from which the earnings call took place (Panel A), and across the type of speaker delivering the sentence (Panel B). Panels A and B show the number of sentences associated to each section of the call, where 'MD (%)' and 'Q&A (%)' show the proportion of call sentences that took place during the Management Discussion and Q&A sections of the calls, respectively. For Panel A, 'Avg Sentiment' shows the arithmetic average sentence sentiment score, whereby the closer the score is to +1 (-1) the more positive (negative) the sentiment, with a sentiment score of 0 corresponding to the neutral sentiment case. 'Avg Audio' shows the average recording length, in seconds (s), for each sentence. For Panel B, 'Mgmt (%)' indicated the proportion of sentences delivered by firm managers on the call, and 'Analyst (%)' shows the proportion delivered by analysts.

Table 1 provides a breakdown of the extracted 2,106-sentence sample across years (Panel A) and speaker type (Panel B). Panel A shows that the number of sentences associated with each earnings call year, disaggregated by call section: Management Discussion (MD) versus Question-and-Answer (Q&A). Average sentiment is reported as the arithmetic mean sentiment score for each year, where scores closer to +1 indicate more positive sentiment, scores closer to -1 indicate more negative sentiment, and scores around 0 correspond to neutral sentiment. Average recording lengths in seconds per sentence are also reported.

The predominance of the Q&A section in Table 1 is evident. The Q&A section of conference calls represent the greatest proportion of these information sharing sessions, within which analysts get the opportunity to question management after the managerial discussion section. The random selection process we use therefore by its very nature extracts more sentences from the Q&A information. This, of course, is more suitable for our analysis given that the analyst-manager interactions tend to be more informative and like to drive sentiment, as has been established in previous literature (e.g., McKay Price et al., 2012; Mayew and Venkatachalam, 2012). Furthermore, from a technical perspective, the sophisticated text-audio alignment process and paralinguistic analysis had much greater levels of success with the Q&A content than the managerial discussion content. This may be due to the unscripted, more animated conversations and interactions between analysts and managers. Management discussions on the other hand tend to be more scripted in nature.

As we do not deliberately stratify by year, the lower number of sentences featured in 2005, 2006 and 2021 is an artifact of the random sampling. Furthermore, average sentiment seems to follow the observed bull and bear trends of the markets. In particular, average sentiment is negative during the

credit crisis and its transition into the longer global financial crisis, and is again negative during the US downturn from 2015-2016.

Our sample of text-audio aligned sentences are pre-processed and split into training and validation sets, where pre-processing includes the removal of special characters, and the transformation of audio data using a scaling function<sup>9</sup> that scales all features to within the range zero to one. Following pre-processing, sentences are randomly assigned to training and validation sets using a stratified train-validation split of 80% training and 20% validation data.<sup>10</sup>

### 2.3. *Paralinguistic Data*

A central component of our approach is the incorporation of paralinguistic data within a sentiment classifier tailored to financial disclosures. The audio features used for the purposes of training and testing our classifier are a subset of audio features provided by the phonetics library *Praat*, namely: mean pitch, mean intensity, number of periods, fraction of unvoiced, number of voice breaks, jitter local, shimmer local, mean autocorrelation, mean noise-to-harmonics ratio and audio length. A definition for each feature is provided in Table 4. The features represent a subset of a larger set of audio features output by *Praat* that were selected based on multicollinearity tests for each variable: features were only included in our classifier if they were not characterised by strong correlations with other variables, to reduce the amount of noise introduced to the multimodal model.

Existing research suggests that the content of what we say matters, and that the way in which we communicate matters (Guyer, Fabrigar and Vaughan-Johnston, 2018). More succinctly, how we speak conveys substantial information beyond the content of communication. Indeed, there is a body of psychology literature that relates to vocal characteristics and their impact on persuasion and/or decision making. For example, evidence suggests that vocal pitch can impact upon listeners' perception of speakers' personal traits and qualities. Several studies associate lower pitch with perceptions of credibility, calmness and maturity, while higher pitch has been associated with nervousness or diminished authority (Song et al., 2020; Chau et al., 2020; Wang et al., 2018; Martín-Santana et al., 2015; Klofstad, Anderson, and Peters, 2012; Feinberg et al., 2005; Chattopadhyay et al., 2003).

From a decision-making perspective Chua et al., (2020) suggest that higher pitched voices increase risk aversion in the listener whilst a lower pitched voice heightens risk tolerance. In a professional setting, Sorokowski et al., (2019) show that both men and woman demonstrate a tendency to lower their mean pitch in an authoritative context, with this effect more pronounced for women. Gelinas-Chebat et al., (1996) define intonation as the variation of pitch which reflects a voice's melodic

---

<sup>9</sup> *MinMaxScaler* from python's scikit-learn library.

<sup>10</sup> A stratified split preserves the same proportion of sentiment categories across both the training and testing set.

contour, with prior evidence suggesting that intonation patterns may, in some contexts, be perceived as indicating lower confidence, while flatter or falling contours may reflect greater authority or competence (Brooke and Ng, 1986; Wallbott, 1982; Apple et al., 1979).

**Table 4.** Definition of Paralinguistic Features

<b>Feature</b>	<b>Description</b>
Mean Pitch	Quality of sound, governed by the rate of vibrations produced; the degree of highness or lowness of a tone.
Mean Intensity	Acoustic power carried by sound waves per unit area in a direction perpendicular to that area.
No. of Periods	Frequency of vibration cycles per second.
Fraction of Unvoiced	Percentage of an audio segment which is unvoiced
No. of Voice breaks	Number of distances between consecutive pulses that are longer than 1.25, divided by the pitch floor.
Jitter Local	Average absolute difference between consecutive periods, divided by the average period.
Shimmer Local	Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.
Mean Autocorrelation	The mean (over all analysed time points) of the autocorrelation, ranging between 0 (theoretical white noise) and 1 (perfectly periodic signal).
Mean NHR	A 'noise-to-harmonics' ratio between periodic and non-periodic components of speech.
Audio Length	The length of each audio clip in seconds

**Notes:** The above table offers concise definitions of each audio variable included in our multimodal sentiment classifier model. Justification for the inclusion of each variable can be found within Section 2.3.

Higher vocal intensity, which can be defined as signal loudness (Gelinac-Chebat et al., 1996), has been associated with increased perceptions of credibility and trustworthiness in some contexts, and a perception to listeners that the speaker has a greater efficiency in articulating their arguments in comparison to softer spoken speakers (Bradac, Mulac, and House, 1988; Brooke and Ng, 1986; Conley, O'Barr and Lind, 1979; Erickson, Lind, Johnson, and O'Barr, 1978). Van Zant and Berger (2020) study the paralinguistic cues used by speakers in nonverbal persuasion attempts, finding evidence that speakers: (i) spoke at a higher volume (greater intensity); (ii) spoke at a higher pitch; (iii) varied their volume more; and (iv) speak at a faster rate. The authors further suggest that speakers are more persuasive when speaking with greater intensity and more varied volume.

Jitter and shimmer audio features have also received some attention with regards to the impact of both cues on speaker perceptions, with studies incorporating these measures in relation to their characterisation of stress, and finding that jitter and shimmer features diminish during experimentally induced stress (Giddens et al., 2013; Park et al., 2010; Mendoza and Carballo, 1998). Furthermore, both features have found to enhance classification performance in some speech recognition and emotion detection tasks. For example, Li et al., (2007) show that the addition of both features to a classification model results in increased classification accuracy, in comparison to a model that only contained baseline spectral and energy features.

The remaining audio cues have received comparatively little coverage in the existing literature, in terms of speakers' perception and listener influence. In machine learning applications, these additional features have been shown to contribute to improved emotion classification performance in specific model contexts. The "fraction of unvoiced" (and also the number of voice breaks) reflecting the proportion (and number) of pauses within an audio clip, represent commonly used vocal characteristics in emotion classification literature. For example, Morrison, Wang and De Silva (2007) adopt seven vocal characteristics from calls received by call centres to classify the emotion displayed on the calls, which includes a "fraction of unvoiced" variable. The authors adopt multiple classification models to assess which model returns the highest accuracy for the task at hand. After performing baseline classifications with each model, feature selection techniques were applied to assess which audio features were optimal to include. For all models, the fraction of unvoiced variable was included after feature selection, highlighting the variables robust correlation with accurate emotion classification.<sup>11</sup>

Using a sample of 10,000 video clips extracted from social media platforms, Morency, Mihalcea and Doshi (2011) classify the sentiment of each video using the fusion of text, audio and visual data. The authors created a proof of concept which suggests that a multimodal approach is affective in identifying online video sentiment. To do so, the authors draw textual and visual data from the videos alongside two audio features: namely pause duration and pitch. The findings suggest that classifying emotion using a tri-modality approach outperforms each of the three modalities in isolation. Furthermore, Poria, Cambria and Gelbukh (2015) incorporate pauses into their vocal modality data for multimodal classification, showing that the inclusion of vocal modality increases classification accuracy. Indeed, even the worst results obtained using two modalities still outperformed the accuracy achieved by single modality models.

---

<sup>11</sup> The authors strongest result came from a voting classifier model that adopted forward selection, which returned an accuracy of 79.43%.

The remaining variables, autocorrelation and noise-to-harmonics ratio (NHR), are quite commonly used in computer emotion and speech recognition tasks (Lee, Kim and Kang, 2014; Rong, Li and Chen, 2009; Nwe, Foo and De Silva, 2003). The benefits of including said variables are demonstrated by Nohroozi et al., (2019), who use a similar range of paralinguistic variables for a vocal based emotion recognition task. The authors improve on past accuracies using this set of paralinguistic variables, highlighting that their enhanced performance is in part due to their smaller set of features, which lowered model complexity whilst improving computational speed and thus emotion classification.

The literature discussed above accentuates the benefits of including each paralinguistic feature within a sentiment classifier as they have been shown in various studies to carry informative insights surrounding the emotional state of the speaker. We incorporate these paralinguistic cues into our finance-specific classifier to evaluate whether features shown to be predictive in other domains also contribute meaningfully to sentiment classification in financial disclosures.

### **3. Multimodal Sentiment Classifier**

Our multimodal sentiment classifier leverages the FinBERT transformer model that has gained popularity in recent years. Our multimodal classifier, however, also incorporates the audio modality from corporate earnings calls. The inclusion of nonverbal cues in studies of financial decision making remains rare in the financial literature (Mayew and Venkatachalam, 2012) despite comparatively higher use in other academic domains. Early work by Mehrabian and Wiener (1967) and Mehrabian and Ferris (1967) highlights that vocal and facial cues are more influential than text in the interpretation of emotional messages, particularly when verbal and non-verbal signals conflict. Although these findings originate from studies outside of a financial context, they reinforce the broader point that nonverbal cues can convey affective content. Furthermore, existing evidence suggests that a combination of text and audio data improves classification accuracy, and consequently creates a more robust representation of sentiment (Houjeij et al., 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Given that both textual and vocal characteristics of earnings calls have been found to be informative (Mayew and Venkatachalam, 2012), and that NLP literature finds a combination of text and audio to significantly increase classification accuracy, the merging of both modalities represents a novel approach within financial sentiment classification and builds on emerging findings in both finance and NLP.

We input numerical representations created using FinBERT from the textual sentences<sup>12</sup> along with the sentence-level numerical representations of paralinguistic features (see Table 4) into a deep neural network classifier (DNN) to make sentiment predictions. To assist in our model building, the optimal hyperparameters of the DNN were identified using the Random Search approach (available in the ‘Keras Tuner’ Python library). Rather than using a trial-and-error method, the Random Search evaluates multiple configurations of layers and nodes to return the optimal set up for a specific DNN problem. We arrive at an optimal structure that has an input layer of 500 nodes, two hidden layers consisting of 250 and 125 nodes, respectively, and an output layer consisting of three nodes. The parameters of the DNN involve a ReLu activation function in the input layer and hidden layers with a softmax activation function for the output layer. Similarly, our model deploys a he\_uniform kernel initialiser and an adam optimiser to evaluate the accuracy metric. The DNN uses a batch size of 150 epochs.

In order to add a degree of robustness to the reported accuracy of our new multimodal sentiment analysis model, we employ two additional models to capture the information contained within paralinguistic data: an ‘audio only’ classifier and a multimodal FinBERT neural network (NN) classifier. The audio classifier takes the audio attributes incorporated into the multimodal classifier in isolation, inputting them into a neural network. By doing so, we gain an understanding of the extent to which audio features alone are reliable in predicting sentiment. This should allow for interpretation as to whether audio features are more effectively considered on their own, or in combination with the textual modality. The second robustness model, the multimodal FinBERT classifier, uses the same neural network and textual features used for the FinBERT model, but includes the addition of paralinguistic features to ensure that any increase in performance is the result of audio feature inclusion, rather than the use of a DNN.

### ***Benchmark models***

To evaluate the performance of our multimodal sentiment classifier, we benchmark against three classes of models: dictionary-based methods, traditional machine learning (ML) algorithms, and deep learning (DL) architectures.

The dictionary-based approach relies on counting the number of positive and negative words within a sentence based on pre-defined lexicons. We apply both a general dictionary (Harvard-IV-4) and a domain-specific financial dictionary (Loughran and McDonald, 2011) to our dataset. While dictionary methods require lower computational resources and are intuitive to apply, they suffer from notable limitations, including fixed word sentiment orientation and limited coverage of domain-specific

---

<sup>12</sup> These are identical to representations used for the text only FinBERT model explained in Section 3.3.

terminology. To complement this, we employ a Naïve Bayes classifier as a traditional ML benchmark. This probabilistic method estimates the probability of a sentence’s sentiment based on word occurrence patterns, operating under the "bag of words" assumption. Sentences are tokenised and stop words are removed prior to being transformed into numerical feature arrays for model training and validation. Finally, for a deep learning benchmark, we deploy BERT (Devlin et al., 2019), a transformer-based model that generates rich contextualised representations of text. BERT is pretrained on a large general corpus and produces a 768-dimensional feature vector for each sentence. These representations are input into a simple neural network classifier with a ReLU activation function at the input layer and softmax activation at the output, optimised using the Adam optimiser. The benchmark comparisons allow us to position our multimodal classifier relative to widely adopted and state-of-the-art textual sentiment classification approaches.

## 4. Results

### 4.1. Classifier Accuracy

Our analysis first reflects upon the multimodal classifier’s ability to accurately predict sentiment across the full dataset of 2,106 manually classified earnings call sentences. We then disaggregate the testing data to assess the model’s accuracy according to the speaker (managers or analysts) and the section of the call (management discussion versus Q&A), enabling us to analyse the specific parts of an earnings call that may be more or less suitable for different classification approaches. The accuracy (AC) of each model discussed in the previous section will be compared directly using the testing accuracy metric (Renault, 2020), where TP represents the number of true positive classifications, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 5 presents the recorded accuracy for each model. In our setting, accuracy appears to increase with model complexity, with deep learning and transformer-based models outperforming simpler baselines. Of the models tested, the audio only classifier represents the least accurate,<sup>13</sup> which

---

<sup>13</sup> This being said, it can be seen from Appendix C that the audio only classifier demonstrates greater accuracy on the alternative datasets mentioned in the footnotes of Section 2.2. In the case of both alternative datasets, the audio only classifier outperforms both dictionary approaches. This is perhaps due to the imbalanced nature of the alternative datasets towards the neutral sentiment category. We see the audio neural network classifier heavily overtraining towards neutral sentences and thus returning a higher classification accuracy due to the imbalanced nature of the dataset.

may reflect the fact that sentence-level sentiment was labelled using textual rather than audio content. The dictionary approaches, Harvard IV4 and Loughran-McDonald (LM), return the lowest training and validation accuracy amongst textual classifiers, which is consistent with previous evidence suggesting that advanced techniques are more robust at detecting financial sentiment than the commonly-used dictionary approaches (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017).

**Table 5.** Overall Call Classification Accuracy Results

Model	Accuracy	
	Training	Validation
Audio Only	40.08%	34.12%
Harvard IV4	42.99%	42.18%
Loughran-McDonald (2011)	49.23%	47.39%
Naïve Bayes	90.14%	56.40%
BERT	74.76%	65.17%
FinBERT	97.80%	73.46%
Multimodal FinBERT NN	<b>99.76%</b>	74.64%
Multimodal FinBERT DNN	<b>99.76%</b>	<b>74.88%</b>

**Notes:** This table outlines the accuracy for each sentiment classifier when the classifier is used to classify the training dataset of 1,684 sentences (in-sample) and the validation dataset of 422 sentences (out-of-sample). The highest accuracy achieved in each case is signified using bold text.

The ML method, using a Naïve Bayes classifier, performs especially well when classifying the sentiment of training data that it has already seen, (90.14%) however this accuracy level drops considerably when classifying the unseen validation dataset (56.40%), potentially indicating an overfitting issue by the ML model when classifying training data. There is a considerable degree of consistency, in that the greatest accuracies are generated by the four DL methods, with the multimodal FinBERT DNN returning the highest in-sample accuracy (99.76%). This represents outperformance of 1.42% and 9.71% when compared to the single (text) modality FinBERT and BERT methods, respectively. That being said, the performance of all classifiers predictably drops when considering the out-of-sample dataset. Even then, however, the multimodal FinBERT DNN classifier (74.88%) achieves higher accuracy than all other text-based classifiers, ranging from a 1.42% improvement over FinBERT to a 32.70% improvement over the Harvard IV4 Dictionary.

The multimodal FinBERT NN model, which also combines text and audio modalities, outperforms the text-only FinBERT model by 1.18% out-of-sample, suggesting that audio features may enhance sentiment classification accuracy in this context, albeit modestly. This finding supports the

possibility that multimodal methods offer an incrementally refined measure of sentiment, which may be useful in future investigations linking financial disclosure and trading behaviour.

**Table 6.** Classifier Accuracy (Out-of-Sample) Disaggregated by Sentiment Category

Model	Sentiment Category		
	Negative	Neutral	Positive
Audio Only	27.66%	48.57%	23.40%
Harvard IV4	26.95%	35.71%	63.83%
Loughran-McDonald	40.43%	73.57%	28.37%
Naïve Bayes	58.87%	38.57%	71.63%
BERT	53.19%	67.14%	75.18%
FinBERT	71.43%	73.05%	75.89%
Multimodal FinBERT NN	70.21%	<b>75.00%</b>	<b>78.72%</b>
Multimodal FinBERT DNN	<b>75.18%</b>	70.71%	<b>78.72%</b>

**Notes:** This table identifies the validation (out-of-sample) accuracy across the three sentiment categories for each method being compared in this study. It highlights what method is most adept at classifying each individual sentiment category and subsequently diagnoses in what areas model fall short.

Table 6 reports the out-of-sample classification accuracy across each sentiment category in the validation set. Here, we are seeking to establish whether our multimodal classifier is adept at predicting positive, negative, or neutral earnings call sentences, relative to our benchmark text-modality models. In other words, if the improved performance of our multimodal classifier is driven by one sentiment class – for example, positive sentiment – then this may suggest that audio cues offer greater consistency for certain sentiment types, such as positive speech. Our results suggest that Multimodal FinBERT DNN returns the highest accuracy for positive and negative sentiment categories, whereas the highest neutral classification accuracy is achieved by Multimodal FinBERT NN. Taken together, these results indicate that multimodal classifiers perform comparatively well across all sentiment categories, while recent transformer-based models generally outperform traditional dictionary and Naive Bayes approaches in this context.

We also find evidence suggesting that domain-specific models, on average, outperform general purpose ones, with finance-specific models (Multimodal FinBERT NN, Multimodal FinBERT DNN and FinBERT) achieving the highest classification accuracy amongst all classifiers. Furthermore, the Harvard Dictionary returns the lowest accuracy for negative classifications (26.96%), which is consistent with prior findings that general dictionaries misclassify words used within a financial context. For example, Loughran and McDonald (2011) tailor their dictionary approach to the negative sentiment

category,<sup>14</sup> using the rationale that negative sentiment has more influence on trading activity. Our evidence offers support to this approach, as we find that the finance-specific dictionary has a substantially higher negative (40.43%) than positive (28.37%) classification accuracy. Our results also suggest that the finance-specific dictionary is better able to classify negative sentiment than the general dictionary (26.95%). Combined, these results highlight the potential value of domain-specific model design when applying sentiment analysis to financial disclosures.

**Table 7.** Breakdown of Accuracy across different sections and participants of the call

Model	MD	Q&A	Mgmt	Analyst
Harvard IV4	34.48%	42.75%	41.78%	42.64%
Loughran-McDonald	51.72%	47.07%	47.56%	47.21%
Audio Only	37.93%	33.84%	38.67%	28.93%
Naïve Bayes	65.52%	55.73%	58.22%	54.31%
BERT	72.41%	64.63%	66.22%	63.96%
FinBERT	<b>93.10%</b>	72.01%	69.04%	<b>77.33%</b>
Multimodal FinBERT NN	89.66%	73.54%	75.56%	73.60%
Multimodal FinBERT DNN	89.66%	<b>73.79%</b>	<b>76.00%</b>	73.60%

**Notes:** This table identifies the validation accuracy across the different sections and participants on the call for each method being compared in this study. where ‘MD’ and ‘Q&A’ show the accuracy rates for earnings call sentences occurring the Management Discussion and Q&A sections of the calls, respectively. ‘Mgmt’ and ‘Analyst’ show the accuracy rates for sentences spoken by Managers and Analysts on the call.

Table 7 reports classification accuracy of earnings calls sentences, disaggregated firstly by call section (Management Discussion and Q&A), and secondly by the originator of the sentence (Managers and Analysts). Our results offer further insights on the relative strengths and weaknesses of each classifier. The multimodal FinBERT DNN achieves the highest accuracy across all subgroups, including sentences from the Q&A section and those delivered by managers, within our validation dataset. This may indicate that paralinguistic features enhance classifier performance in more conversational contexts, such as those typically found in the Q&A section of earnings calls. This finding aligns with prior evidence suggesting that sentiment expressed during the Q&A section of earnings calls has greater predictive power for market characteristics than sentiment expressed during the management discussion (McKay Price et al., 2012; Borochin et al., 2017; Fu et al., 2019).

#### 4.2. Additional Accuracy Measures

<sup>14</sup> The negative word list created by Loughran and McDonald (2011) has 2,337 words. When compared to their positive list of 353 words, the negative word list is substantially larger and highlights the authors’ focus on the negative sentiment category.

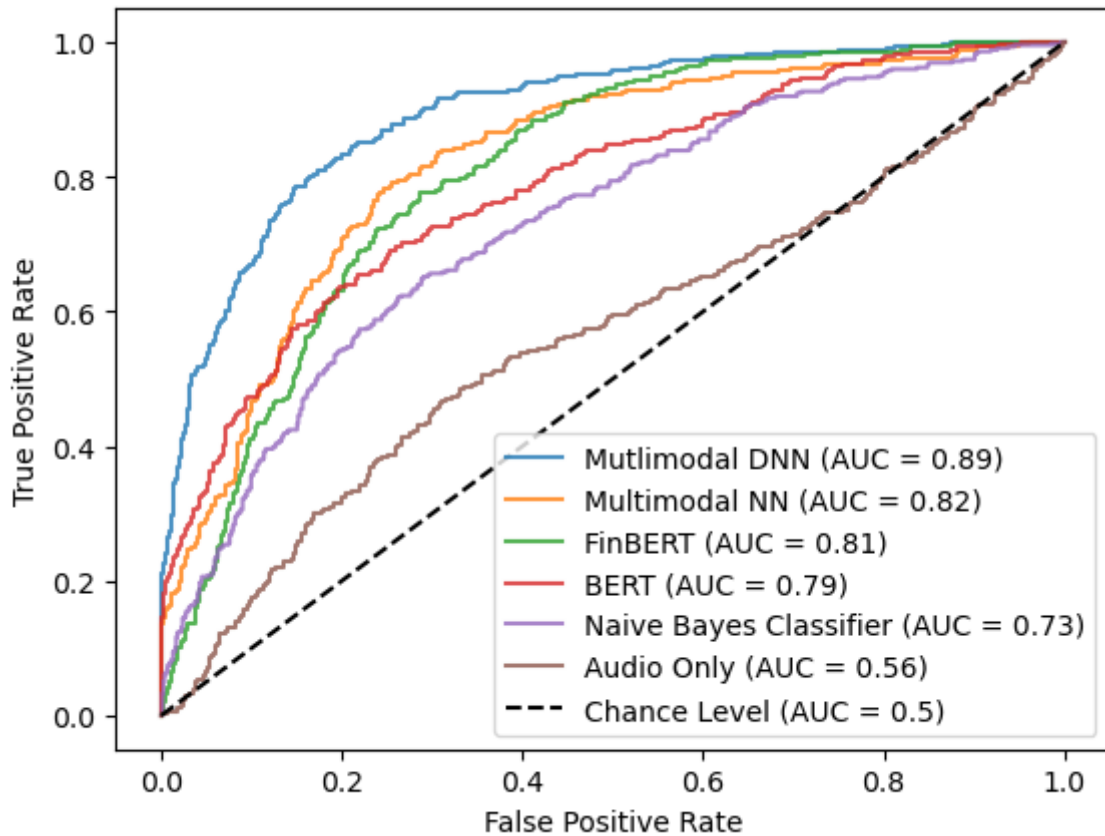
To further test the findings reported in Section 4.1, we evaluate the performance of our multimodal, benchmark and robustness models using Receiver Operating Characteristic curve and Area Under the Curve scores (ROC-AUC).<sup>15</sup> Given that ROC-AUC measures require a binary classification problem, and that we use three classification categories for the purposes of our study (positive, neutral and negative), the methods used to understand each model's ability to classify must be slightly adjusted. Specifically, a 'One versus Rest' (OvR) method, which evaluates each class against all others, was used. The OvR method evaluates each classifier's ability to correctly assign positive sentiment, by assigning a value of one to correct positive classifications, and zero to misclassifications (i.e. neutral or negative sentiment). The OvR model then adopts the same process for the neutral and negative categories. The average accuracy score across all three categories is then recorded for each classifier. In adopting this process, the ROC curve identifies the sensitivity (true positive rate) and specificity (true negative rate) of each sentiment classifier.

Mandrekar (2010) highlights that an ROC curve that intersects the coordinates (0,0) and (1,1) at a 45-degree angle represents pure chance, with any curve skewing towards the upper left corner of the plot representing a classification accuracy above that achieved by chance. Furthermore, the AUC score is an effective way to summarise the overall accuracy of a model, with an overall score of zero indicating a perfectly inaccurate classifier, and a score of one reflecting perfectly correct classification. Results of the ROC-AUC test are shown in Figure 4.

The ROC curve offers provides further evidence supporting the classifier accuracies observed in Section 4.1. Specifically, the ROC curves for each method skew toward the top left corner of the plot, and the AUC scores also gradually increase in the same fashion, as the classifier increases in (i) complexity and (ii) accuracy. The Multimodal FinBERT DNN achieves the highest AUC score (0.89). This is followed by the Multimodal FinBERT NN (0.82), although this performance reflects only a marginal improvement on the single-modality FinBERT NN (0.81). Furthermore, the single-modality audio classifier is characterised by an ROC curve that lies very close to the pure chance line, and an AUC scores (0.56) only marginally higher than the chance level (0.50). Combined, the results suggest that audio characteristics in isolation only offer an incrementally increased ability to accurately predict sentiment.

---

<sup>15</sup> Due to the configuration of Dictionary Methods, it is not possible to create an ROC curve, and hence this method has been omitted from Figure 4.

**Figure 4.** Receiver Operating Characteristic (ROC) curves

**Notes:** This figure plots macro-averaged ROC curves for each of the models used within this study apart from the dictionary-based methods. It identifies the trade-off between True positive rates and False positive rates for each model. It also displays the AUC score for each model which is an effective summary of model accuracy. The closer an AUC score is to 1 the more robust the model is at making correct classifications.

#### 4.3. Paralinguistic Feature Importance

The results reported in Sections 4.1 and 4.2 indicate that incorporating audio features into sentiment analysis models may enhance classification accuracy in the context of financial disclosure. However, it may be the case that this enhanced performance is being driven by a small number of informative paralinguistic features, in which case, those audio cues that are barely informative may be removed from more streamlined dual-modality models in future. For this reason, we employ permutation importance analysis to assess the extent to which each of the specific paralinguistic features established in Table 4 inform the multimodal FinBERT DNN model, with results presented in Table 8.<sup>16</sup>

<sup>16</sup> We calculate permutation importance using Python's *eli5* library.

**Table 8.** Feature Importance Weights for Paralinguistic Features

<b>Audio Feature</b>	<b>Weight</b>
Fraction of Unvoiced	$0.0811 \pm 0.0168$
Shimmer Local	$0.0622 \pm 0.0152$
Mean Pitch	$0.0443 \pm 0.0113$
Mean NHR	$0.0442 \pm 0.0126$
Mean Autocorrelation	$0.0406 \pm 0.0118$
Number of Periods	$0.0368 \pm 0.0158$
Audio Length	$0.0337 \pm 0.0185$
Number of Voice Breaks	$0.0230 \pm 0.0071$
Jitter Local	$0.0229 \pm 0.0144$
Mean Intensity	$0.0126 \pm 0.0052$

**Notes:** This table represents the feature importance of each paralinguistic feature within the multimodal DNN model. The weight column represents the weight of each feature in relation to the other paralinguistic features in the dataset. The number to the left of the  $\pm$  is the mean weight estimate with the number to the right of the symbol being the standard deviation of the estimate.

Permutation importance evaluates the importance of each feature in a classification model by measuring the impact of each feature on model accuracy, when specific features are randomly shuffled. If the model accuracy decreases with the shuffled feature data, the feature is considered to be important as it no longer carries the same level of information. Our results suggest that all audio variables contribute to the multimodal DNN model’s predictions to varying degrees, which is somewhat expected given that our paralinguistic features were selected from a larger list of features, based on their prevalent use in prior studies outside of the finance domain.<sup>17</sup>

However, it is evident from our results that some features are more important than others. Namely, the fraction of unvoiced feature is found to be almost twice as informative as the third most informative feature (mean pitch). Shimmer local and the noise-to-harmonics ratio are also shown to be particularly important. Our findings are generally consistent with prior literature suggesting that the fraction of unvoiced (Morrison, Wang and De Silva, 2007), shimmer (Li et al., 2007; Jacob, 2016) and pitch (Koolagudi and Rao, 2010; Koolagudi, and Krothapalli, 2012; Chebbi and Jebara, 2018) variables are important features for sentiment/emotion classification. Mean intensity is found to be least informative, though it does still hold a very small degree of influence on the predictions made by the multimodal FinBERT DNN classifier.

## 6. Conclusion

<sup>17</sup> The method used to select the final set of features is provided in Section 2.1.

Textual analysis methods, namely sentiment analysis, have become increasingly popular within the academic finance literature in recent years. The techniques used to determine sentiment in each case vary considerably, with the most popular approaches – such as dictionaries and Naïve Bayesian classifiers – being comparatively more rudimentary than recent advancements, such as transformer architecture. Our study offers two main contributions. Firstly, we build on the comparative work of Frankel et al., (2022) by creating a contemporary comparison of the most used sentiment analysis methods in academic finance, alongside approaches commonly adopted in other domains. To do so, we use a dataset of 2,106 audio-text aligned sentences extracted from corporate earnings calls relating to twenty constituents of the S&P 100 index. Our results suggest that transformer-based and deep learning models outperform more traditional methods in classifying sentiment from earnings call content, at least within our dataset. Secondly, we show that the addition of a second modality, in the form of paralinguistic features, improves classification accuracy relative to text-based models, in the context of our earnings call sample of S&P100 firms.

Our findings are consistent with the previous literature highlighting computationally advanced approaches to be more robust at capturing financial sentiment than commonly used dictionary methods (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Munikar, Shakya and Shrestha, 2019; Sun et al., 2020; Alamoudi and Alghamdi, 2021). Furthermore, the findings accentuate the conclusions made by Mayew and Venkatalcham (2012), and the wider social psychology literature, by finding that non-verbal information is incrementally informative in the communication process, albeit the effect is small. In a financial context, our results suggest that paralinguistic cues provide incremental insight into the sentiment conveyed in corporate earnings calls.

Although our multimodal sentiment classification model returns the highest classification accuracy for this dataset, there is still scope for further enhancement, mainly in regard to the creation of the classified sentiment data. For example, though our classifier is domain-specific to finance, it is not industry- or firm-specific. The companies used for the purposes of our study operate in a number of different industries and settings that differ, both in terms of the terminology used and the complexity of business models. Incorporation of these factors, perhaps through the use of industry experts to manually classify the training set, and validate results, could potentially improve the contextual understanding of the sentiment classifiers used.

Additionally, future work could involve dual-modality manual labelling, incorporating both text and audio, in order to better align with the multimodal nature of the inputs and potentially enhance the contribution of paralinguistic features. Finally, a manually classified dataset would allow deeper evaluation of model performance, particularly given the known scalability of deep learning models with data volume. Despite these limitations, our multimodal model demonstrates improved classification

accuracy relative to common approaches in the finance literature, and may provide a useful foundation for future research.

## References

Abirami, A.M. and Gayathri, V., 2016. A survey on sentiment analysis methods and approach. *2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)*. Chennai, India: Institute of Electrical and Electronics Engineers, pp. 72–76.

Alamoudi, E. and Alghamdi, N., 2021. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 30(2-3), pp.259-281.

Antweiler, W. and Frank, M., 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, 59(3), pp.1259-1294.

Apple, W., Streeter, L.A. and Krauss, R.M., 1979. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5), pp. 715–727.

Bhaskar, J., Sruthi, K. and Nedungadi, P., 2014. Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*.

Bhonde, R., Bhagwat, B., Ingulkar, S. and Pande, A., 2015. Sentiment Analysis Based on Dictionary Approach. *International Journal of Emerging Engineering Research and Technology*, 3(1), pp. 51–55.

Borochin, P., Cicon, J., DeLisle, R. and Price, S., 2017. The effects of conference call tones on market perceptions of value uncertainty. *Journal of Financial Markets*, 40, pp.75-91.

Bradac, J.J., Mulac, A. and House, A., 1988. Lexical diversity and magnitude of convergent versus divergent style shifting-: Perceptual and evaluative consequences. *Language & Communication*, 8(3-4), pp. 213–228.

Brockman, P., Li, X. and Price, S., 2015. Differences in Conference Call Tones: Managers vs. Analysts. *Financial Analysts Journal*, 71(4), pp.24-42.

Brooke, M.E. and Ng, S.H., 1986. Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, 5(3), pp. 201–210.

Chattopadhyay, A., Dahl, D., Ritchie, R. and Shahin, K., 2003. Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology*, 13(3), pp. 198–204.

Chebbi, S. and Ben Jebara, S., 2018. On the use of pitch-based features for fear emotion detection from speech. *2018 4<sup>th</sup> International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* [Preprint].

Chua, G.Y.P., Er, H., Liaw, S. and He, T., 2020. Pitch Right: The Effect of Vocal Pitch on Risk Aversion. *Economics Bulletin*, 40(4), pp. 3131–3139.

Conley, J.M., O’Barr, W.M. and Lind, E.A., 1979. The power of language: Presentational style in the courtroom. *Duke Law Journal*, 1978(6), p. 1375.

D’Andrea, A., Ferri, F., Grifoni, P. and Guzzo, T., 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3), pp. 26–33.

- Dair, Z., Donovan, R. and O'Reilly, R., 2021. Classification of Emotive Expression Using Verbal and Non-Verbal Components of Speech. *2021 32nd Irish Signals and Systems Conference (ISSC)*.
- Devlin, J., Chang, MW., Lee, K. and Toutanova, K., 2019. Bert: Pre-training of deep bidirectional Transformers for language understanding. [online] arXiv.org.
- Dey, L., Chakraborty, S., Biswas, A. and Bose, B., 2016. Sentiment analysis of review datasets using naïve Bayes' and k-nn classifier. *International Journal of Information Engineering and Electronic Business*, 8(4), pp. 54–62.
- Diesner, J. and Evans, C., 2015. Little Bad Concerns. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*.
- El-Haj, M., Rayson, P., Walker, M., Young, S. and Simaki, V., 2019. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, 46(3-4), 265-306
- Erickson, B., Lind, EA., Johnson, BC. and O'Barr, WM., 1978. Speech style and Impression Formation in a court setting: The effects of 'powerful' and 'powerless' speech. *Journal of Experimental Social Psychology*, 14(3), pp. 266–279.
- Feinberg, D.R., Jones, BC., Little, AC., Burt, DM. and Perrett, DI., 2005. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, 69(3), pp. 561–568.
- Frankel, R., Jennings, J. and Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, 68(7), pp.5514-5532.
- Fu, X., Wu, X. and Zhang, Z., 2019. The Information Role of Earnings Conference Call Tone: Evidence from Stock Price Crash Risk. *Journal of Business Ethics* 173, 643–660.
- Gélinas-Chebat, C., Chebat, J.-C. and Vaninsky, A., 1996. Voice and advertising: Effects of intonation and intensity of voice on source credibility, attitudes toward the advertised service and the intent to buy. *Perceptual and Motor Skills*, 83(1), pp. 243–262.
- Giddens, C.L., Barron, K., Byrd-Craven, J., Clark, K. and Winter, A., 2013. Vocal indices of stress: A Review. *Journal of Voice*, 27(3).
- González-Bailón, S. and Paltoglou, G., 2015. Signals of Public Opinion in Online Communication. *The ANNALS of the American Academy of Political and Social Science*, 659(1), pp.95-107.
- Grimmer, J. and Stewart, B., 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), pp.267-297.
- Guo, L., Shi, F. and Tu, J., 2016. Textual analysis and machine learning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, 2(3), pp.153-170.
- Guyer, J.J., Fabrigar, L. and Vaughan-Johnston, T., 2018. The counterintuitive influence of vocal affect on the efficacy of affectively-based persuasive messages. *Journal of Experimental Social Psychology*, 74, pp. 161–173.
- Houjeij, A., Hamieh, L., Mehdi, N. and Hajj, H., 2012. A novel approach for emotion classification based on fusion of text and speech. *2012 19th International Conference on Telecommunications (ICT)*.
- Jacob, A., 2016. International Conference on Communication and Signal Processing. *Speech emotion recognition based on minimal voice quality features*.
- Kearney, C. and Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, pp.171-185.

Klofstad, C.A., Anderson, R.C. and Peters, S., 2012. Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738), pp. 2698–2704.

Koolagudi, S.G. and Rao, K.S., 2010. Real life emotion classification using VOP and pitch based spectral features. *2010 Annual IEEE India Conference (INDICON)* [Preprint].

Koolagudi, S.G. and Krothapalli, S.R., 2012. Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. *International Journal of Speech Technology*, 15(4), pp. 495–511.

Lee, J.W., Kim, S. and Kang, H.-G., 2014. Detecting pathological speech using contour modeling of harmonic-to-noise ratio. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [Preprint]

Li, F., 2010. The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, 48(5), pp.1049-1102.

Li, J., Yang, L., Smyth, B. and Dong, R., 2020. MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 3063–3070.

Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K. and Newman, J., 2007. Stress and emotion classification using Jitter and Shimmer Features. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* [Preprint].

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: *International Conference on Learning Representations*.

Loughran, T. and McDonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35-65.

Loughran, T. and McDonald, B., 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4), pp.1187-1230.

Mandrekar, J.N., 2010. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9), pp. 1315–1316.

Martín-Santana, J.D., Muela-Molina, C., Reinares-Lara, E. and Rodriguez-Guerra, M., 2015. Effectiveness of radio spokesperson's gender, vocal pitch and accent and the use of music in radio advertising. *BRQ Business Research Quarterly*, 18(3), pp. 143–160.

Mayew, W. and Venkatachalam, M., 2012. The Power of Voice: Managerial Affective States and Future Firm Performance. *The Journal of Finance*, 67(1), pp.1-43.

McKay Price, S., Doran, J., Peterson, D. and Bliss, B., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4), pp. 992-1011.

Mehrabian, A., and Ferris, S. R., 1967. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3), 248–252.

Mehrabian, A., and Wiener, M., 1967. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6(1), 109–114.

Mendoza, E. and Carballo, G., 1998. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12(3), pp. 263–273.

Morency, L.-P., Mihalcea, R. and Doshi, P., 2011. Towards multimodal sentiment analysis. *Proceedings of the 13th international conference on multimodal interfaces* [Preprint].

- Morrison, D., Wang, R. and De Silva, L.C., 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2), pp. 98–112.
- Munikaar, M., Shakya, S. and Shrestha, A., 2019. Fine-grained Sentiment Classification using BERT. *2019 Artificial Intelligence for Transforming Business and Society (AITB)*.
- Nogueira, R. and Cho, K., 2020. Passage Re-ranking with BERT. *Cornell University Working Paper*.
- Nwe, T.L., Foo, S.W. and De Silva, L.C. 2003. Speech emotion recognition using Hidden Markov models. *Speech Communication*, 41(4), pp. 603–623.
- Park, C.-K., Lee, S., Park, H., Baik, Y., Park, Y.B. and Park, Y.J., 2010. Autonomic function, voice, and Mood States. *Clinical Autonomic Research*, 21(2), pp. 103–110.
- Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing – EMNLP '02* [Preprint].
- Poria, S., Gelbukh, A. and Cambria, E., 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal: Association for Computational Linguistics.*, pp. 2539–2544.
- Renault, T., 2017. Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, 84, pp.25-40.
- Renault, T., 2020. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, 2(1-2), pp.1-13.
- Ribeiro, F., Araújo, M., Gonçalves, P., André Gonçalves, M. and Benevenuto, F., 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1).
- Rong, J., Li, G. and Chen, Y.-P.P., 2009. Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*, 45(3), pp. 315–328.
- Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. and Pantic, M., 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65, pp. 3–14.
- Song, S., Baba, J., Nakanishi, J. and Yoshikawa, Y., 2020. Mind The Voice!: Effect of Robot Voice Pitch, Robot Voice Gender, and User Gender on User Perception of Teleoperated Robots. *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery*, pp. 1–8.
- Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B., and Pisanski, K., 2019. Voice of authority: Professionals lower their vocal frequencies when giving expert advice. *Journal of Nonverbal Behaviour*, 43(2), pp.257–269.
- Sprenger, T., Tumasjan, A., Sandner, P. and Welpe, I., 2013. Tweets and Trades: The Information Content of Stock Microblogs. *European Financial Management*, 20(5), pp. 926-957.
- Sun, C., Qiu, X., Xu, Y. and Huang, X., 2020. How to Fine-Tune BERT for Text Classification? *Chinese Computational Linguistics*, pp.194-206.
- Todd, A., Bowden, J., Moshfeghi, Y., 2023. Text-Based Sentiment Analysis in Finance: Synthesizing the Existing Literature and Exploring Future Directions. *Intelligent Systems in Accounting, Finance and Management (Forthcoming)*.
- Troussas, C., Espinosa, K., Llaguno, K. and Caro, J., 2013. Sentiment analysis of Facebook statuses using naive Bayes classifier for language learning. *IISA 2013* [Preprint].

Van Zant, A.B. and Berger, J., 2020. How the voice persuades. *Journal of Personality and Social Psychology*, 118(4), pp. 661–682.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I., 2017. *Attention Is All You Need*. [online] arXiv.org.

Wallbott, H.G., 1982. Contributions of the German? Expression psychology? to nonverbal communication research. *Journal of Nonverbal Behavior*, 7(1), pp. 20–32.

Wang, T.-Y., Kawaguchi, I. and Kuzuoka, H., 2018. Effect of Manipulated Amplitude and Frequency of Human Voice on Dominance and Persuasiveness in Audio Conferences. *Proceedings of the ACM on Human-Computer Interaction*. New York, NY: Association for Computing Machinery.

Yang, Y., Mark Christopher, M. and Huang, A., 2020. *Finbert: A pretrained language model for Financial Communications*, [online] arXiv.org.

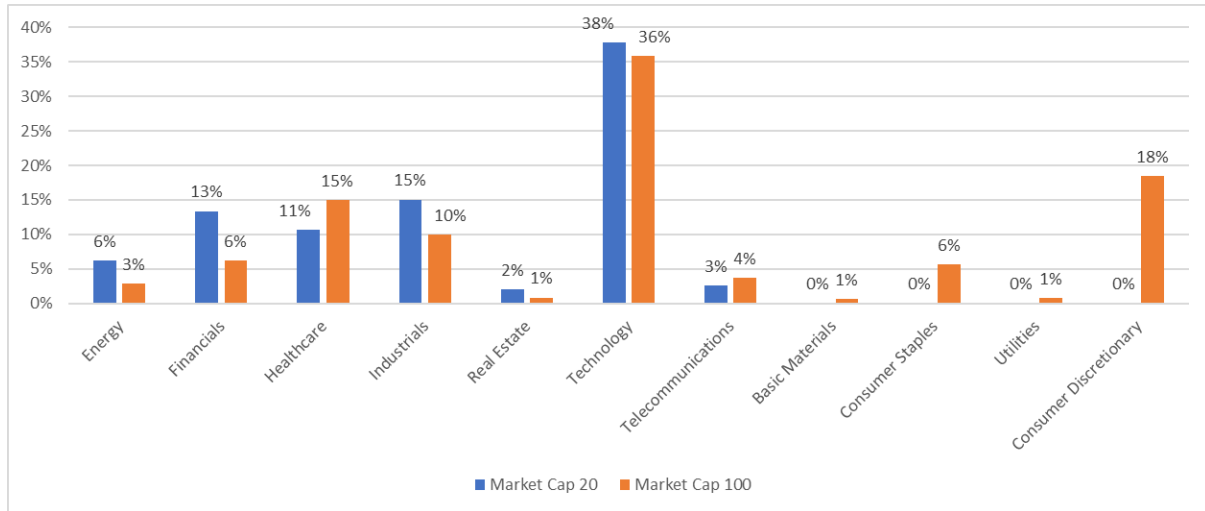
Yan, K., Xu, H. and Gao, K., 2020. CM-BERT. *Proceedings of the 28th ACM International Conference on Multimedia*.

## Appendix

### Appendix 1. Sample of Twenty S&P100 Constituent Firms

<b>Ticker</b>	<b>Market Cap (% of Sample)</b>	<b>ICB Industry name</b>	<b>No. of Calls</b>
ADBE.O	4.24%	Technology	61
AXP	2.49%	Industrials	57
BLK	2.29%	Financials	52
CSCO.O	3.88%	Telecommunications	63
CVX	6.17%	Energy	57
GOOG.O	32.10%	Technology	64
GS	2.44%	Financials	55
HON.O	2.74%	Industrials	58
JPM	7.23%	Financials	54
LLY	6.00%	Health Care	58
LMT	2.33%	Industrials	54
LRCX.O	1.42%	Technology	59
MA	6.99%	Industrials	57
MRK	4.61%	Health Care	59
ORCL.K	4.24%	Technology	61
PLD	2.06%	Real Estate	54
PNC	1.43%	Financials	53
SYK	1.71%	Health Care	55
T	2.63%	Telecommunications	61
UNP	2.99%	Industrials	54
<b>Total</b>	<b>100.00%</b>		<b>1146</b>

**Appendix 2. Comparison of Industry Weighting: Our Sample of Twenty Companies versus the S&P100 Index**



**Notes:** This table compares the market capitalisation of our sample of 20 companies against the wider S&P100. The market capitalisation percentages in blue represent each industries market share for the sub index of 20 firms with the market capitalisations in orange representing each industries overall market share in the S&P100 index.

Journal Pre-proof

## A Multimodal Sentiment Classifier for Financial Decision Making

### Highlights:

- We develop a deep learning multimodal sentiment classifier combining text and paralinguistic audio features from earnings calls.
- We construct a large sentence-level dataset aligning audio and text from twenty S&P 100 firms between 2005 and 2021.
- Our multimodal model outperforms dictionary, Naïve Bayes, and transformer-based text-only models in out-of-sample accuracy.
- Findings highlight the incremental value of audio cues for financial sentiment analysis, particularly in Q&A earnings call sections.

Journal Pre-proof