Received: 30 September 2022 | Revised: 7 December 2023 | Accepted: 29 January 2024

DOI: 10.1002/isaf.1549

REVIEW ARTICLE

WILEY

Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions

Andrew Todd[1] | James Bowden[1] | Yashar Moshfeghi[2]

[1]Strathclyde Business School, University of Strathclyde, Glasgow, UK

[2]NeuraSearch Laboratory, University of Strathclyde, Glasgow, UK

Correspondence
Andrew Todd, Strathclyde Business School, University of Strathclyde, 199 Cathedral St, Glasgow G4 0QU, UK.
Email: andrew.todd@strath.ac.uk

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

Summary

Advances in Deep Learning have drastically improved the abilities of Natural Language Processing (NLP) research, creating new state-of-the-art benchmarks. Two research streams at the forefront of NLP analysis are transformer architecture and multimodal analysis. This paper critically evaluates the extant literature applying sentiment analysis techniques to the financial domain. We classify the financial sentiment analysis literature according to the most used techniques in the area, with a focus on methods used to detect sentiment within corporate earnings conference calls, because of their dual modality (text-audio) nature. We find that the financial literature follows a similar path to NLP sentiment literature, in that more advanced techniques to define sentiment are being used as the field progresses. However, techniques used to determine financial sentiment currently fall behind state-of-the-art techniques used within NLP. Two future directions stem from this paper. Firstly, we propose that the adoption of transformer architecture to create robust representations of textual data could enhance sentiment analysis in academic finance. Secondly, the adoption of multimodal classifiers in finance represents a new, currently underexplored area of study that offers opportunities for finance research.

KEYWORDS
earnings conference calls, multimodal sentiment classifier, natural language processing, natural language understanding, sentiment analysis, text analysis, transformer architecture

1 | INTRODUCTION

Since the arrival of the internet on a commercial scale in the mid-1990s, the manner in which information is delivered to investors, and how investors respond to such information, has been altered considerably (Nardo et al., 2016). Specifically, an increase in the amount of available digital information, facilitated by the scale of interactions that can now be documented, has led to a situation whereby the vast amount of unstructured data that investors have access to—in the form of corporate disclosures, news media, email and social media—often make rational and perfectly informed decisions unattainable (Chan & Chong, 2017). Thus, text mining applications developed within the Computer Science literature have played a key role in processing and condensing large, unstructured textual datasets into data that can be more easily absorbed by time-constrained investors with limited cognitive abilities.

Evidence suggests that investors react to public news (Chan, 2003). Natural Language Processing (NLP) techniques have therefore become increasingly popular in the finance literature as a method of deriving quantitative representations of the sentiment conveyed within a range of financially relevant texts, such as corporate disclosures (Jiang et al., 2019; Loughran & McDonald, 2011), earnings call transcripts (Brockman et al., 2015; Chen et al., 2018), newspaper articles (Bowden et al., 2019; Garcia, 2012; Tetlock, 2007) and social media (Bollen et al., 2011; Gu & Kurov, 2020; Renault, 2020).

Kearney and Liu (2014) provide a comprehensive overview of textual sentiment methods and models within academic finance, and the amount of relevant research has grown substantially in the years following publication. Further, advances in computational power and the recent development of state-of-the-art sentiment analysis methods, such as GloVe and BERT, have given rise to additional sentiment analysis methods within the academic literature in recent years (Daudert, 2021).

Therefore, the purpose of this paper is to comprehensively survey the extant literature which has applied NLP techniques to financial information sources, with a focus on sentiment analysis techniques. In doing so, we seek to (i) ascertain the extent to which sentiment analysis can reveal additional information to the marketplace, (ii) identify trends—in terms of methods and models used—within the recent literature over time, and (iii) investigate future applications and extensions of text-based sentiment analysis, specifically with regards to the use of multimodal sentiment classifiers that incorporate audio, as well as textual, cues.[1]

Through surveying the relevant literature, this paper highlights that there is potential for state-of-the-art methods to be used in detecting and classifying financial sentiment. For example, the application of models that adopt transformer architecture to produce more reliable sentiment measures is a relatively new endeavour which represents an exciting future direction. However, doubts and concerns over such methods must be overcome, with questions regarding the reproducibility and replicability of previous studies utilising machine learning methods (Jiang, 2021). Cambria and White (2014) highlight that the continuous search for more accurate approaches is because of the automatic analysis of text involving a deep understanding of natural language by machines, a reality we have yet to reach.

We also highlight the potential additional information gained through the application of more than one modality of information (multimodal sentiment classifiers) to achieve greater classification accuracy and identify new associations between news and investor reactions. Specifically, we focus primarily on earnings conference calls as an exciting avenue for research in this area as it is a financial disclosure that offers multiple modalities (text and audio) and because of the potential asymmetry-reducing effects of managerial disclosure (both intentional and unintentional). Further, suppose managerial tone is found to have some impact on the price discovery process, as detected by NLP techniques. In that case, this may provide a foundational basis for researchers to incorporate additional modalities beyond textual analysis, namely paralinguistic features, given the audio-based nature of earnings call interactions, and early application of audio analysis techniques within the finance domain (Mayew & Venkatachalam, 2012).

The remainder of this paper is structured as follows. Section 2 focusses on the earliest approaches to textual analysis in finance, such as general and finance-specific dictionary approaches, before progressing through the literature towards more computationally demanding approaches employed in recent research. Section 3 then specifically evaluates earnings calls as a medium for sentiment analysis and proposes several future avenues for research in both textual analysis and multimodal analysis. Section 4 offers some concluding remarks.

2 | APPLICATIONS OF SENTIMENT ANALYSIS IN FINANCE

2.1 | General dictionary approach

Sentiment analysis is the computational study of people's opinions, attitudes and emotions towards an entity (Medhat et al., 2014). Many of the earliest studies employing sentiment analysis techniques within accounting and finance utilise a dictionary approach, also known as the 'word count' approach (Guo et al., 2016; Loughran & McDonald, 2016). The concept behind this approach is comparatively intuitive compared to more recent machine learning methods, in that the sentiment conveyed within a financial text is determined by a count of words within the text that also appear within pre-defined word lists[2] (Li, 2010). Expanding on this, Bhonde et al. (2015) note that word lists are created by first collecting a set of general sentiment words with known positive or negative implications. Once this initial list is created, it is then expanded upon by including synonyms and antonyms for the sentiment words. This iterative process of expanding the word lists ends when no new words can be found. After the process of collecting synonyms and antonyms ends, an inspection of the words is usually completed to clean up the lists.

Abirami and Gayathri (2016) highlight that the most basic way to define sentiment using these dictionaries is to count the number of positive and negative words within a body of text with reference to the dictionary categories. After this count is complete, a comparison of how many positive versus negative words in the text infers how positive or negative the text is. Utilisation of dictionary methods presents advantages in comparison to machine learning techniques; mainly, less computational power (or resource) is required to create and use the dictionaries. However, there are also considerable drawbacks to this approach. For example, the lexicons are characterised by a finite number of words and the sentiment orientation for each word is fixed, resulting in a lack of accuracy in context or domain-specific classification (Abirami & Gayathri, 2016; Bhonde et al., 2015; D'Andrea et al., 2019).

The most popular general word lists used within existing finance research are the Harvard IV psychosocial word lists[3] (Kearney & Liu, 2014), from which Stone and Hunt (1963) created the General Inquirer (GI) system, which leveraged these general-purpose word lists for content analysis in the domain of social psychology. The GI has since been frequently utilised in academic finance (see Tetlock, 2007; Tetlock et al., 2008; Twedt & Rees, 2012). Tetlock (2007) uses the Harvard word lists to assess the impact of news sentiment[4] on the Dow Jones Industrial Average (DJIA) and Standard & Poor's 500 (S&P) indices and finds that a one standard deviation change in the level of pessimism expressed in financial news[5] drives an 8.1 basis point change in DJIA returns. Expanding on this research, Tetlock et al. (2008) use a similar approach and dataset to identify that a one standard deviation increase in negative words translates into a 3.2 basis point reduction in next day abnormal returns. Twedt and Rees (2012) apply the GI to financial analyst reports and show that a change from the lowest quartile of analyst report tone (most pessimistic) to the highest quartile of analyst report tone (most optimistic) results in an average increase in return of 0.7%, holding all else equal. It is, however, worth noting that none of the aforementioned studies employing the GI method identify a profitable trading strategy when factoring in transaction costs.

Davis and Tama-Sweet (2012) utilise another commonly used general dictionary, DICTION,[6] to assess the differences in managers' language across regular earnings press releases (EPRs) and annual 10K statements. The authors find that, on average, 1.08% (1.01%) of the words in 10K filings are optimistic (pessimistic) in nature, compared to 1.27% (0.46%) of words in EPRs. This implies that EPRs convey higher levels of optimism and lower levels of pessimism, in respect to annual filings. Davis and Tama-Sweet (2012) note that information contained in EPRs is processed more efficiently than that contained within 10Ks.[7] Together, these findings suggest that managers potentially anticipate stronger market reactions following EPRs and thus strategically adopt an optimistic tone in these disclosures.

Bollen et al. (2011) and Siganos et al. (2014) each use general dictionaries to test for relationships between the sentiment conveyed within social media posts and index returns. Incorporating Google Profile of Mood States (GPOMS) and OpinionFinder (OF) tools,[8] Bollen et al. (2011) employ a self-organising Fuzzy Neural Network to predict the next day's change in DJIA index values, based upon the three previous days. The authors observe a classification accuracy of 73.3%. However, when considering the calm sentiment indicator from GPOMS in addition to the previous prices, the accuracy increases to 86.7%. This evidence suggests that sentiment indicators can be robust in increasing market value forecasting. Siganos et al. (2014) adopt Facebook's Gross National Happiness Index[9] to assess Facebook sentiment's relationship with returns, trading volume and volatility across 20 international markets.[10] They demonstrate that an increase of 0.1 in the sentiment measure translates into a 31-basis point increase in international market returns.[11]

2.2 | Domain-specific dictionary approach

Loughran and McDonald (2011) demonstrate that general dictionaries misclassify words used within a financial context, noting that 73.8% of negative words within the Harvard dictionary are not considered negative in a financial context. González-Bailón and Paltoglou (2015) and Ribeiro et al. (2016) also demonstrate the limitations of general dictionaries in classifying content in domain-specific settings. Both authors do this by applying a general dictionary to text stemming from various domains and show that the reliability and validity across these differing sets are low. One alternative to this issue is to create domain-specific dictionaries, where adding words to an existing dictionary and deleting irrelevant words (or words with different meanings) within a specific context would be beneficial (Diesner & Evans, 2015; Grimmer & Stewart, 2013).

To the authors' knowledge, Henry (2006) is the first to use a finance-specific word dictionary to overcome the domain-specificity limitation[12] inherent in general dictionaries (Chan et al., 2021). Specifically, the author creates positive and negative word lists through the inspection of past EPRs. Using a word count approach, the author evaluates the extent to which sentiment can be used to improve accuracy in forecasting S&P500 index returns. Whereas a model only using financial variables returns a forecasting accuracy of 54.12%, the accuracy increases to 59.52% when including the sentiment measure. Thus, forecasting ability, when incorporating the sentiment conveyed within earnings releases, is found to increase by 5.4%. A later study by Henry (2008) lends support to these findings through the identification that greater levels of positive tone within corporate press releases result in higher abnormal returns even after controlling for financial results.[13] Furthermore, the market reaction increases with the level of positive tone conveyed, up until a certain point.[14]

To overcome the issue of domain-specific terminology, Loughran and McDonald (2011) also build financial dictionaries which include categories relating to negative, positive, uncertain, litigious, strong modal and weak modal words. The authors highlight that their primary focus is the negative dictionary.[15] To create these word lists, the authors developed dictionaries of all words and their word counts relating to the above categories stemming from all 10Ks filed from 1994 to 2008. They then carefully examined all words that occurred in at least 5% of all documents and created final word lists based upon the top 5% most used terms in the financial documents. These word lists (hereafter referred to as the LM dictionary) have been widely used throughout the literature for word count sentiment analysis approaches (Bannier et al., 2017; Ferguson et al., 2015; Garcia, 2012; Jegadeesh & Wu, 2012; Jiang et al., 2019; Johnman et al., 2018; Mao et al., 2011).

Jiang et al. (2019) leverage the LM word lists to create a manager sentiment index to forecast future aggregated S&P 500 index market returns[16] and find that a one standard deviation increase in sentiment relates to a 1.26 standard deviation decrease in S&P 500 returns. Furthermore, a high manager sentiment is associated with low excess aggregate market returns in the next month, suggesting that overvaluation occurs when the manager sentiment index is high, leading to low future stock returns.[17]

Garcia (2012) and Ferguson et al. (2015) both use the LM dictionary to assess media article sentiments relationship with DJIA and FTSE100 returns, respectively. Garcia (2012) creates a pessimism factor—calculated by subtracting the number of positive words within a text from the number of negative words—from New York Times media articles and evaluates the impact of pessimism on returns over recessionary and expansionary periods.[18] For the expansionary period, a one standard deviation change in the pessimism factor precedes a market movement of 3.5 basis points of DJIA returns. In relation to the recessionary period, a one standard deviation increase in the pessimism factor is associated with a 12-basis point increase in the DJIA. All tests return significant results, indicating that sentiment helps predict next day stock returns.[19] Ferguson et al. (2015) find similar results when analysing media articles[20] relevant to the UK market; both positive and negative sentiments conveyed in UK news predict returns on the same day as the publication. Specifically, a one standard deviation increase in positive (negative) words increases (decreases) abnormal returns by 4.9 (2.3) basis points. Curiously, the authors show that the significant predictive relationship between media and next period abnormal returns is driven by less visible firms. Thus, highly visible firms within the FTSE100 experience less pronounced effects from positive and negative words in news stories. However, Johnman et al. (2018) find that sentiment has no significant relationship with daily excess returns for firms within the FTSE100 and, through the creation of trading-based strategies including transaction costs, find no economic value in trading based on news sentiment.[21]

Bannier et al. (2017) analyse the performance of the German Deutscher Aktien Index (DAX) in reaction to sentiment (defined using the LM dictionary) conveyed in CEO speeches given during firm AGMs. Changes in negative and positive sentiment are found to be strongly associated with cumulative abnormal returns (CARs), calculated from the day before the AGM to 30 days following. Specifically, an increase in negative (positive) sentiment of 0.749 (0.353) corresponds with a decrease (increase) in CARs of 2.77% (3.14%). However, when considering the immediate market reaction,[22] it is found that negative sentiment has no significant relationship with abnormal returns, whereas positive sentiment has a small association in economic terms.

Mao, Counts and Bollen et al. (2011) create a negative news sentiment (NNS) indicator by applying Loughran and McDonald's (2011) negative word lexicon applied to financial news headlines, to evaluate the sentiment measures in relation to the DJIA market index. They find that the NNS is significantly correlated to market log returns (−0.147). Furthermore, they find statistically significant Granger causation in both directions between log returns and the NNS.

2.3 | Machine learning approaches

The studies mentioned in this review thus far provide strong evidence that sentiment defined using the specific dictionary approach captures a more accurate measure of market response to earnings calls than its general counterpart.[23] In recent years, advances in computational power have allowed for the application of Machine Learning (ML) methods for the purposes of sentiment analysis within finance. Multiple papers have compared the accuracy of dictionary methods (both general and domain-specific) with the ML approach. For example, McGurk et al. (2020), Renault (2017) and Guo et al. (2016), all provide evidence that ML approaches are more accurate at classifying financial sentiment. However, Renault (2020) states that ML models may be sufficient in deriving textual sentiment from online sources but warns that more complex algorithms do not necessarily equate to more accurate results.

A commonly used ML algorithm to detect sentiment is the probabilistic Naive Bayesian classifier, which considers the naïve independence assumption[24] and is commonly used as a baseline method for classifying text (Dey et al., 2016). Naïve Bayes methods estimate the probability that a document is positive or negative given its contents. It estimates the probability of a word being 'positive' or 'negative' in nature by looking through a series of positive and negative texts and counting how often the word appears in each (Troussas et al., 2013). Hence, a crucial part of this model's method is pre-classified data to train on. Dey et al. (2016), however, note a benefit of the Naïve Bayes classifier is that it only requires a small amount of training data to establish parameters necessary for classification.

Antweiler and Frank (2004) present the first study to the author's knowledge to adopt an ML approach to classifying financial sentiment. The authors employ a Naïve Bayes classifier to evaluate the relationship between internet financial message board interactions and both the DJIA Index and Dow Jones Internet Commerce Index (XLK), finding that a one standard deviation increase in bullishness[25] translates into a 1.75 standard deviation increase in abnormal returns. Using similar methods, Li (2010) analyses the extent to which sentiment conveyed within forward-looking statements from the Management Discussion and Answer (MD&A) section of 10Ks and 10Qs is associated with contemporaneous abnormal returns, finding that a one standard deviation increase in sentiment relates to a 4.7 basis point increase in returns.

Sprenger et al. (2013)) also use a Naïve Bayes classifier to determine sentiment conveyed within Twitter interactions and test for associations with S&P 100 index returns. The authors identify a statistically strong but economically weak relationship between bullish sentiment and index returns. Specifically, a one standard deviation increase in bullish sentiment of tweets is associated with a 0.5 basis point increase in returns. However, the authors show that their sentiment measures cannot be used to predict returns, whereas the effect of returns on sentiment is positive and significant. Hence, returns affect sentiment but not vice-versa.

Groß-Klußmann and Hautsch (2011) adopted the Reuters NewsScope Sentiment Engine (RNSE) to retrieve 29,497 news headlines with accompanying sentiment and relevance indicators.[26] The findings suggest that the machine-indicated relevance of news is supported by market reactions. In other words, there is a significantly stronger reaction to the news if the news has been ranked with high relevance.[27] Evaluating the difference in reaction to initial news and subsequent updates, the authors find that trading on updated news is much more pronounced than trading on initial news. These findings support the notion of news clustering[28] and suggest that the reiteration and reinforcement of news create stronger signals, which translate into stronger market reactions.

Audrino and Tetereva (2019) evaluate sentiment spill-over effects, focussing specifically on whether news sentiment (defined using RNSE) has cross-industry effects for the S&P 500 and the Euro Stoxx 50 indexes. The authors note that the relevance of news stemming from differing sectors shows fluctuating effects on returns that are spread evenly among industries. However, there is evidence of finance and energy news holding a greater influence across all sectors.

These influential sectors have spill-over effects that seem to be at least as important as the direct effects of their sentiment.[29]

Adopting similar methods, Sun et al. (2016) use the RNSE to evaluate sentiment at the intraday level for S&P 500 index returns. In this case, the authors use a combined dataset of sentiment conveyed within financial news, SEC filings, social media and earnings calls. Sun et al. (2016) show evidence that their lagged sentiment measure (split into half-hour periods across the day) is a robust predictor of last half-hour intraday returns. A one standard deviation increase in sentiment results in a 0.269 standard deviation increase in returns in the last half hour of the trading date.

The impact of Twitter sentiment on asset prices is further investigated by Azar and Lo (2016). The researchers focus on a dataset of Tweets that mention terms that are related to the Federal Open Markets Committee (FOMC), such as ‘FOMC’, ‘Federal Reserve’, ‘Bernanke’ or ‘Yellen’ on the basis that decisions made by the FOMC are popular among the investment community and significantly affect asset prices (Bernanke & Kuttner, 2005; Cieslak et al., 2014; Lucca & Moench, 2015). The authors find that tweet sentiment can be used to predict day-ahead returns, with the effect intensifying on days when the FOMC meet.[30] A one standard deviation increase in tweet sentiment on FOMC days results in an increase of 0.58% in returns the following day. Interestingly, tweet sentiment on days that the FOMC do not meet becomes negligible. A trading strategy based on tweet sentiment on days in which the FOMC meet is found to passively track the CRSP value-weighted index on all except for 8 days a year (when the FOMC meets) and significantly outperforms the market benchmark over a 1-year period.

Gu and Kurov (2020) adopt Bloomberg's Twitter sentiment measure to assess its ability to forecast returns for the US-focussed Russell 3000 index,[31] and find that the sentiment measure has a statistically significant contemporaneous correlation of 0.14 on average with index returns. On average, the stock return over the following 1-day period for firms with the most positive social media sentiment is roughly 27.2 basis points higher than the return for firms subject to the most negative sentiment. Curiously, the authors demonstrate that the coefficient estimate for sentiment is much smaller for the equal-weighted index (0.048) in comparison to that of the value-weighted index (0.136). These findings suggest that sentiment does have more predictive power for the returns of small firms relative to large firms. Finally, Gu and Kurov (2020) create two portfolios at the start of each trading day, going long on firms with high positive sentiment and short on firms with negative sentiment, before rebalancing at the beginning of every day. Ignoring transaction costs, this strategy is found to return a daily average of 8.6 basis points, which translates into a 21.5% annual return with a Sharpe ratio of 3.17.

As discussed, financial disclosures have been subject to a steady stream of sentiment analysis literature within recent years. Figure 1 illustrates this trend over time by providing an overview of the number of published studies utilising different sentiment analysis methods, disaggregated by publication year for all articles referenced within this review, with finance-specific dictionaries and machine learning methods gaining in popularity in recent years.

2.4 | State-of-the-art natural language processing approaches

The previous sections of this paper discuss the application of dictionary and machine learning approaches to the financial domain and

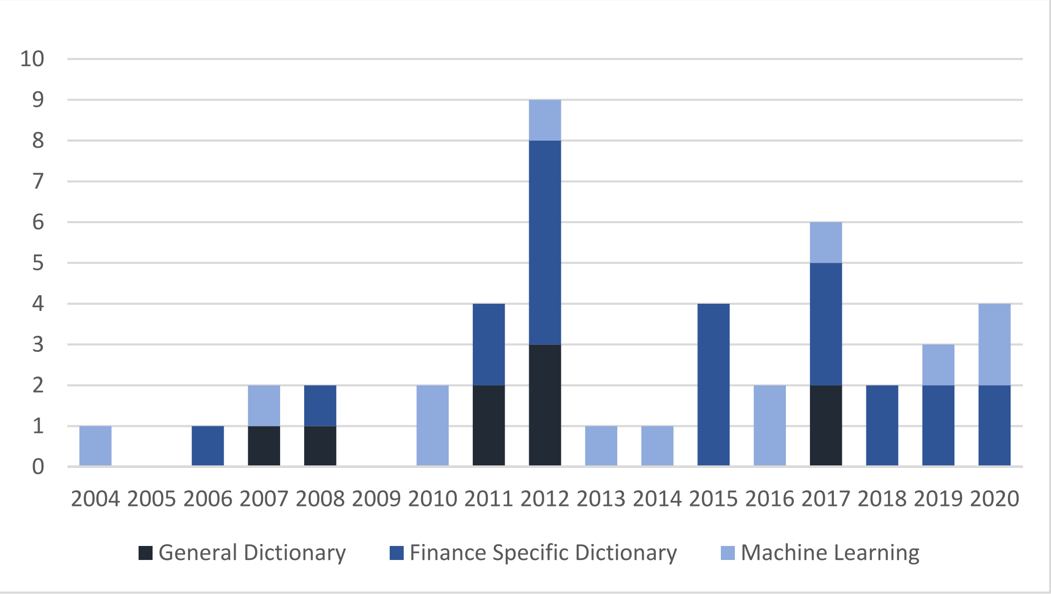

FIGURE 1 Frequency of published studies applying sentiment analysis methods to financial data. *Notes*: This figure shows the number of published studies within academic finance that have utilised sentiment analysis techniques to investigate associations between financial sentiment and trading activity. The annual frequency is broken down into three categories, based on the specific technique used to derive sentiment.

highlight that as the techniques used to define financial sentiment increase in complexity, so too does the accuracy of the captured sentiment. However, El-Haj et al. (2019) identify that the field of accounting and finance falls behind that of NLP studies in the classification of sentiment using the state-of-the-art methods. They note that there is a scarcity of advanced NLP techniques being applied in the financial domain.[32] While the ML techniques discussed in the previous section have been shown to classify financial sentiment better than more rudimentary approaches, alternative approaches such as transformer architecture (Alamoudi & Alghamdi, 2021; Munikar et al., 2019; Sun et al., 2019) and multimodal analysis (Bhaskar et al., 2014; Dair et al., 2021; Houjeij et al., 2012; Yang et al., 2020) have been demonstrated as having greater abilities in accurately capturing sentiment.

2.4.1 | Transformer architecture

Before the introduction of the transformer by Vaswani et al. (2017), the authors highlighted that state-of-the-art results across various NLP tasks were dominated by Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) (Chung et al., 2014; Hochreiter & Schmidhuber, 1997). Instead of attempting to push state-of-the-art results by improving on previous RNN language models or Encoder–Decoder architecture (Jozefowicz et al., 2016; Luong et al., 2015; Wu et al., 2016), Vaswani et al. (2017) introduced the transformer, which is a model that is solely based on attention mechanisms and does not require recurrence or convolutions. Because of the complexities of transformer architecture, we do not provide a comprehensive overview of model architecture within this paper. However, due to the common implementation of this computation model, multiple in-depth descriptions are available, for example, Vaswani et al. (2017).[33]

Many models since the introduction of transformer architecture have returned state-of-the-art results in various tasks by adopting and building upon the initial method. For example, Raffel et al. (2020) introduced a text-to-text transfer transformer (T5), which has achieved state-of-the-art performance on the SQuAD question and answering task.[34] Brown et al. (2020) train an autoregressive language model (GPT3) on 175 billion parameters,[35] which returned the highest accuracy of 86.4% on the LAMBADA language modelling task.[36] A third model which effectively utilises the transformer is Bidirectional Encoder Representations from Transformers (BERT). Devlin et al. (2019) introduce BERT and compare it to various other advanced models across multiple datasets. The authors show that the general BERT model, not pretrained on any specific data only finetuned towards the specific tasks, performed competitively (80.5% accuracy, representing a 7.7% absolute improvement on GLUE).[37]

Across the three models discussed above, BERT performs particularly well on sentiment classification tasks (Alamoudi & Alghamdi, 2021; Munikar et al., 2019; Sun et al., 2019). However, the only paper to the authors knowledge to use BERT in the financial domain is Hiew et al. (2019), who applies BERT to posts on the Chinese social media platform Weibo relating to three listed firms on the Hong Kong Stock Exchange (HKSE)—Tencent, Ping An and CCB. The author compares their method with commonly used machine learning methods within the established literature,[38] finding that BERT vastly outperforms each of the comparison models on three key criteria.[39] These findings support the suggestion that BERT demonstrates stronger capabilities of financial sentiment classification in the Chinese language over common ML models.

Similar to traditional dictionary approaches, Howard and Ruder (2018) show that transformer model performance for text classification can be significantly improved when further pretrained on a domain-specific corpus. Huang et al. (2023) created a financial version of BERT named FinBERT that is pretrained on 4.9 billion tokens from three financial corpora: earnings conference call transcripts, annual reports and analyst reports. They compare FinBERT to BERT across three different financial sentiment analysis tasks; Financial Phrase Bank,[40] AnalystTone[41] and FiQA.[42] Finding that the model pretrained on general language, BERT, does not perform as well as the FinBERT model pretrained on financial language.

Transformer architecture is pushing the capabilities of machines in understanding text. The studies cited within this section indicate an enhanced performance in NLP tasks and a more adept understanding of the nuances of textual communication. This deeper understanding of the communication process holds numerous future avenues for research of qualitative financial data, particularly with regards to the ability to gain a deeper understanding of the impact and function of qualitative data in financial markets. However, as with all methods discussed previously in this review, it has limitations. Khan et al. (2022) provide a survey of the application of transformers for computer vision. They highlight the main limitations of transformers being high computational cost because of their size and complexity, large data requirements because of their need for a substantial amount of good quality data for training and the models' poor interpretability because of their complex architectures.

2.4.2 | Multimodal analysis

Most sentiment analysis techniques, across all subject fields of academic research, have mainly used singular modality-based models—in the most part, text-based classifiers, which have been shown to be useful for various tasks such as forecasting box office revenues (Asur & Huberman, 2010), election outcome prediction (Tumasjan et al., 2010), classifying customer reviews (Gräbner et al., 2012) and—as the aforementioned literature suggests—stock market prediction.

Audio and visual data have also been used singularly to identify sentiment in recent years. The analysis of speech for emotion classification has been researched extensively (Koolagudi & Rao, 2012). More recently, studies have evaluated the ability of vocal cues to define sentiment alone. Multiple papers have shown that audio cues can successfully define sentiment (Pereira et al., 2014; Kaushik et al., 2013; Mairesse et al., 2012; Mayew & Venkatachalam, 2012). Soleymani et al. (2017) note that the field of sentiment analysis using visual data alone has not been fully researched. Indeed, the sole research to the author's knowledge in this area is Borth et al. (2013).

Soleymani et al. (2017) highlight that recent developments in this branch of natural language understanding have started to consider the combination of modalities.[43] Multimodal sentiment analysis can be defined as the inclusion of additional modalities (audio and/or visual) to compliment text-based models in an attempt to improve sentiment classification. Extending the input data of sentiment classifiers is gaining traction because of its usefulness in assessing sentiment on a plethora of publicly accessible multimodal data platforms such as Facebook, YouTube, Reddit, and Twitter (Gandhi et al., 2023). The gold standard for multimodal sentiment analysis is considered to be the combination of all three communication modalities—text, audio and visual. Various studies have used the combination of all three modalities to define sentiment, showing that the use of a tri-modality model is more robust at classifying sentiment over bi-modal and singular modality models (Bhaskar et al., 2014; Dair et al., 2021; Houjeij et al., 2012; Morency et al., 2011; Poria et al., 2015; Yang et al., 2020).[44]

The main advantage of using multimodal classifiers for sentiment classification is the additional behavioural cues provided by the visual and audio data.[45] The insights that vocal and visual data provide are substantial and allow for a more robust sentiment to be captured. However, there are limitations of multimodal sentiment analysis, particularly in its application to the financial domain. The limitations come in the form of access to multimodal data, adhesion of the different modalities into a successful classifier, and the generalisation of multimodal models. The application of multimodal analysis in finance poses issues due to the lack of data sources that contain more than one modality—the only dual modality reliable data source in finance to the authors' knowledge is earnings conference calls.[46]

Gandhi et al. (2023) highlight that there are various methods to fuse together the different modalities in a multimodal classifier. In their review of sentiment literature, it is evident that there is a substantial lack of analysis of text–audio multimodal classifiers and subsequently a lack of consensus on the best way to extract and fuse together these two modalities—this is evidently an obstacle when evaluating earnings calls as the two modalities stemming from said disclosure are text and audio. Finally, multimodal sentiment analysis models do not generalise well. If a model has been trained on a specific person/or set of people and learned their behavioural cues, the results of the model on another set of people do not scale to true generalisation.

3 | APPLICATION OF SENTIMENT ANALYSIS ON EARNINGS CONFERENCE CALLS

> ‘Tesla Inc. investors gave a rare rebuke to iconoclastic Chief Executive Elon Musk on Wednesday after he cut off analysts asking about profit potential, sending shares down 5 percent despite promises that production of the troubled Model 3 electric car was on track.’ Reuters, 2nd May 2018

Corporate earnings calls allow for additional qualitative or ‘soft’ information to be revealed by managers to the marketplace (Blau et al., 2015). This increase in the amount of information available to financial analysts serves as a useful mechanism to reduce informational asymmetries between the firm and the investment community (Bowen et al., 2002). Unlike company disclosures, which are prepared and reviewed in advance of publication, unexpected questions posed by analysts in real time present an opportunity for managers to reveal information that they had not planned to disclose, or—in the case of the Reuters article quoted above—react in a way that the market interprets negatively. As a result, the earnings call mechanism has been subject to a steady stream of sentiment analysis literature in recent years.

Earnings calls are a channel of communication whereby company managers, commonly Chief Executive Officers (CEOs) and Chief Financial Officers (CFOs), provide a statement surrounding past, present and future firm performance and answer questions posed by interested parties (such as analysts, institutional investors and individual investors). Frankel et al. (1999) suggest that the contents of conference calls provide additional information to the market and document the rapid growth of firms utilising earnings calls, to the extent that some 92% of companies represented by the National Investor Relations Institute (NIRI) actively run earnings calls. McKay Price et al. (2012), Doran et al. (2012) and Matsumoto et al. (2011) provide support for this statement, finding that earnings call participants are actively engaged to the extent that new and meaningful information comes to light. This additional information is the product of analysts and institutional investors' continued participation and probing for information, alongside the supplementary insights managers sometimes provide above that contained within the press release. Earnings calls are structured in a different format compared to other qualitative data communications used within the industry, and these differences potentially allow for new information to be unearthed.

In recent studies, the distinct setup of earnings calls—particularly the two differing sections[47] and the nature of participants on the call—has provided an opportunity for varied research. Authors have produced research evaluating associations between specific managers (Davis et al., 2015; Davis & Tama-Sweet, 2012; Larcker & Zakolyukina, 2012; Mayew & Venkatachalam, 2012) or analysts (Milian & Smith, 2017) sentiment and returns. Comparisons between manager and analyst sentiment (Borochin et al., 2017; Brockman et al., 2015; Chen et al., 2018), and the actions of investors in response (Amoozegar et al., 2020; Blau et al., 2015; Bochkay et al., 2020; Mayew & Venkatachalam, 2012), have also been examined. However, research in this area primarily focusses on the overall sentiment of a call—sentiment calculated and aggregated based on all call participants (Borochin et al., 2017; Doran et al., 2012; Fu et al., 2019; McKay Price et al., 2012; Wang & Hua, 2014).

3.1 | Overall tone

To the author's knowledge, McKay Price et al. (2012) conducted the first investigation of associations between the sentiment of earnings calls and securities pricing. Controlling for the numerical representation of the earnings surprise, the authors demonstrate that positive and negative earnings call sentiment—defined using the Henry (2006) finance-specific dictionary—is significantly related to (i) abnormal returns during the initial earnings announcement window[48]; (ii) the post-earnings announcement drift; and (iii) abnormal trading volume. Further, the researchers find that qualitative information in the form of earnings calls has greater explanatory power on subsequent returns over longer time horizons,[49] in comparison to actual earnings figures. More succinctly, the market may find it easier to incorporate numerical data, but qualitative data provided in the earnings call format are shown to provide additional value relevant information that is not so rapidly incorporated. Particularly, the Q&A section of the call holds significant ability in predicting CARs, post earnings drift and abnormal trading volume, when controlling for numerical earnings surprise and the sentiment of the prepared remarks. Thus, earnings calls, representing the only financial disclosure to contain natural language conversations surrounding firm performance, present a rich source of information.

Focussing specifically on Real Estate Investment Trusts (REITs),[50] Doran et al. (2012) measure the extent to which linguistic sentiment[51] produced in earnings calls is associated with future fluctuations in market value. Consistent with McKay Price et al. (2012), earnings call sentiment is found to have significant explanatory power over abnormal returns at the market level. Interestingly, the authors note that firms whose earnings calls contain substantial positive (negative) sentiment have higher (lower) abnormal returns. Furthermore, their analysis produces findings that indicate positive call sentiment can completely offset negative earnings surprises for low earnings surprise firms.[52] These results suggest that managers have the potential to improve firm performance by using positive linguistic terminology during calls.

Borochin et al. (2017) also identify earnings calls as an important medium for disseminating information to the market. However, unlike previous studies, the authors focus on uncertainty rather than abnormal returns.[53] The results indicate that higher levels of pessimism lead to greater pricing uncertainty, with higher levels of optimism creating the opposite effect.[54] The authors separate earnings call sentiment into three distinct aspects: (i) the manager's sentiment during the call introduction; (ii) the manager's sentiment during the Q&A session; and (iii) the analyst sentiment during the Q&A. The authors find that both managers and analysts impact upon investor uncertainty with differing magnitudes. Negative coefficients for the managerial introductory element of the call were identified,[55] suggesting that value perceptions of investors for the upcoming quarter are slightly impacted by a managers' introductory statement. However, no meaningful relationship is established for manager Q&A sentiment, thus implying their sentiment within this section of the call is less influential. In comparison to managers, the expression of negative analyst Q&A sentiment is shown to heavily influence investor uncertainty.[56] This suggests that analysts are perhaps viewed in a more trusted and objective light on the call by market participants. Hence, analyst sentiment receives more market attention and thus holds a potentially greater influence on share prices.

Most recently, Fu et al. (2019) analysed associations between earnings call sentiment[57] and stock price crash risk,[58] finding that higher levels of optimism on quarterly calls negatively predict stock price crash risk with statistical and economical significance.[59] Thus, higher optimism reduces stock price crash risk. This conclusion is somewhat anticipated given that optimism is typically associated with positively performing companies, and stock price crash risk is associated with struggling institutions. Consistent with McKay Price et al. (2012), the Q&A section of the call is also found to have greater predictive power over market pricing than the introduction section, reinforcing that the market pays closer attention to the Q&A section of the call.

However, there is a lack of consensus as to the most informational aspect of earnings calls. For example, Fu et al. (2019) provide conflicting evidence suggesting that managers' sentiment throughout the Q&A section has stronger and more statistical prediction power. This conflicts with Borochin et al. (2017) who conclude that 'managers, as corporate insiders, possess private information and engage in truthful communication during conference calls'. Thus, indicates that managers generally remain truthful on conference calls and do not attempt to mislead participants to improve their performance, even in the face of extreme downside risk.

3.2 | Managerial/analyst sentiment and styles

A recurring theme throughout the above papers is the impact that manager-specific sentiment and style has on the market reaction to earnings calls. The first paper to look in-depth at managerial sentiment is Larcker and Zakolyukina (2012), who evaluate whether the language of executives can assist in unearthing financial reporting manipulation or misstatements. In their analysis of individuals occupying managerial positions, they find that deceptive CEOs and CFOs have distinct traits in common. For example, both (i) use more references to general knowledge; (ii) use fewer non-extreme positive words; and (iii) limit their discussions surrounding shareholder value. These findings infer that the consideration of managerial linguistic features offers a valuable tool in understanding the quality of financial reporting. Finally, the researchers assert that linguistic models applied in this setting dominate, or are at least equate to, models that are based on purely accounting and financial information.

Mayew and Venkatachalam (2012) evaluate nonverbal communication on earnings calls but, unlike Larcker and Zakolyukina (2012), focus on managerial affective states[60] in relation to future firm performance. The findings suggest that a manager's vocal cues allow analysts to learn about a manager's affective state, and in turn about the firm's financial future. Supporting this, the authors find that a one standard deviation increase in positive affective state, defined by vocal cues, is positively related to unexpected earnings of 6.9 and 7.53 basis points unexpected earnings over the next two to three periods.[61] Similarly, a one standard deviation increase in negative affective states decreases unexpected earnings by 3.07 and 4.31 basis points over the same period. To the author's knowledge, Mayew and Venkatachalam (2012) are the first to study nonverbal communication in a capital market setting, and thus provide the foundations for numerous subsequent studies employing this different modality. It is, however, an area still in its relative infancy.

In the same vein, Davis et al. (2012) assesses the effect that managers other than the chief executive have on the sentiment of earnings calls and seeks to identify the extent to which manager-specific optimism impacts the language used in firms' conference calls. Managers' sentiment and language choice are found to be strongly and positively related to market reaction. Further evidence of the ability for managerial optimism to influence the market response is demonstrated through an increase in adjusted R-squared from 9.95% in the base regression (including no manager variables) to 10.6% with the addition of managerial fixed effects.[62] Finally, the overall tone of the call is influenced by the arrival of an optimistic or pessimistic manager to the firm, thus suggesting that market reactions can be impacted by individual managers. These findings are perhaps particularly important as they suggest that managerial sentiment on a conference call does not only reflect the private knowledge a manager has surrounding his/her firm but that it is also a product of manager-specific tendencies towards optimism or pessimism. The authors show multiple factors which are associated with these tendencies, for example, manager age, volunteerism, work experience and gender.

Building on earlier research, Davis et al. (2015) evaluate (i) whether each manager has a specific, consistent sentiment,[63] (ii) whether this sentiment remains constant across different firms they work for and, (iii) if the sentiment can be measured. It is found that manager-specific style can be detected and measured, and that it does stay constant even when the manager moves to a new firm. Further, observable managerial characteristics are identified, which play a strong role in explaining the sentiment outputs generated. For instance, managers who started their career in a recession use less positive language, along with managers who have investment banking experience. However, managers with charitable involvement tend to be more positive.

Evidently, managerial sentiment and style have been shown as informative variables for the analysis of relationships between earnings calls and financial market activity. However, specific research into analyst attitudes on calls remains understudied in the existing literature. Milian and Smith (2017) provide an exploratory analysis by creating a list of compliments used commonly by analysts on earnings calls, through an extensive reading of call transcripts. With this corpus at their disposal, they investigated analyst compliments, denoted as praise, on earnings calls to unearth potential relationships with the market. The underlying logic for this argument is thus analysts may complement managers to curry favour and build relationships, to gain a better position in accessing private information. Alternatively, they may compliment in an unbiased manner based upon the merits disclosed in the earnings announcement. The authors find that praise is significantly and positively associated with abnormal earnings announcement stock returns. Specifically, a one standard deviation increase in praise coincides with a positive abnormal return of 1.34%. In comparison to traditional sentiment measures,[64] the authors find that their ‘praise’ dictionary is 3.5 times stronger in predicting the magnitude of abnormal returns. Furthermore, these results indicate that praise given by analysts is given accordingly, and not excessively produced when not merited, thus dismissing the idea that analysts compliment managers to curry favour. Curiously, praise is found to be statistically significant and related to future stock returns while both sentiment measures are not (overall sentiment and analyst sentiment). Thus, analyst compliments, defined as praise, look to be a robust variable in understanding positive firm performance.

3.3 | Comparison of manager versus analyst sentiment

An evident gap in the literature on earnings call characteristics and financial market activity concerns the comparison of managerial sentiment to analyst sentiment. Brockman et al. (2015) provide a rare foray into this area, finding that managers speak with significantly greater optimism, at greater length and use less complex language in comparison to their call counterparts. This is somewhat expected, given that managers are disseminating information and in doing so want to communicate clearly and in a positive manner surrounding their firm, whereas analysts attend to gain further information surrounding earnings figures and future performance.

Furthermore, the authors demonstrate that at the time of the call, managerial and analyst sentiment is significantly associated with stock prices,[65] and overall positive (negative) sentiments are related to positive (negative) abnormal returns. Finally, the results indicate that both managerial and analyst sentiment are quickly incorporated into stock prices, with analyst sentiment gaining a stronger market reaction, suggesting that market participants lend more credence to variance in analyst sentiment in comparison to that of managerial sentiment. This lends further weight to the suggestion that the difference between both parties is not trivial.[66]

In related research, Chen et al. (2018) analyse manager–analyst conversations on earnings calls. Their research evaluates the potential impact that manager–analyst sentiment has on intraday stock prices.[67] In their exploration of communication exchange, these authors identify a similar theme to many prior papers: analyst sentiment carries more weight in market reactions. Based on the aforementioned evidence that price fluctuations are larger over the interactive section of the call, the authors seek to establish which features of a call drive such events. The results suggest that intraday stock prices significantly respond to analyst sentiment with evidence suggesting that the effect strengthens when analyst sentiment is relatively negative.[68]

Two accompanying findings may help to understand this relationship. Firstly, in comparison to management, analysts speak in a more neutral fashion. Secondly, both participants' sentiment moves away from an optimistic tone as the call progresses. Thus, if analysts speak in a neutral fashion and are not biased, changes in sentiment thus reveal further information surrounding performance. Furthermore, call sentiment is initially optimistic because of managerial introductions and, throughout the duration of the call, begins to move towards a sentiment which fits with the firm's performance of the previous quarter. Analysts seem to begin the call closer to this level of sentiment; thus, investors utilising analyst sentiment may identify the informational content of the call quicker. Combined, these findings point strongly in favour of analyst sentiment being more influential in the market setting, and thus more useful from a commercial perspective.[69]

3.4 | Investor response to sentiment

There exists a lack of consensus in the literature with regard to the extent to which investors respond to earnings call sentiment. Blau et al. (2015) evaluate the extent to which short sellers incorporate 'soft' qualitative information (p.203) into their forecasts, thus allowing them to understand if and how investors gauge and use sentiment. In doing so, they look to understand the extent to which detected abnormal sentiment[70] is acted upon by short sellers. The findings lend support to the suggestion that short sellers do use soft information from earnings calls when valuing their stocks and do trade against firms with positive earnings surprises and high abnormal sentiment.

Most recently, Bochkay et al. (2020) evaluated the impact of extreme words on market digestion.[71] Creating a corpus of extreme words, the authors first find that abnormal returns are much more strongly correlated to extreme language in relation to moderate words. A one standard deviation increase in extreme language results in a 6.9% increase in abnormal trading volume. They further show that over a 60-day period, there is no inclination of reversals or price drifts. This implies that investors price extreme language in earnings calls correctly. Furthermore, the authors use analyst revisions to evaluate analyst reactions to extreme language. The findings suggest that, over a 10-day period, language extremity is strongly associated with analyst revisions. Furthermore, analysts react more strongly to extreme positive language. Both findings are consistent with the idea that sentiment is an effective measure of market characteristics and shows that the incorporation of sentiment measures is being used by the commercial sector.

4 | CONCLUSIONS

When evaluating the wealth of academic studies utilising sentiment analysis techniques within academic finance, a clearly defined trend emerges: techniques used to define sentiment are increasing in complexity to capture the most robust sentiment measures possible. This is perhaps no surprise, given that relationships between market variables and sentiment calculated using more advanced sentiment classification techniques are shown to be stronger in comparison to sentiment defined using more fundamental approaches (Kearney & Liu, 2014; McGurk et al., 2020; Renault, 2017). Though the number of studies employing advanced techniques has been increasing in recent years, the bulk of the literature to date has been conducted with comparatively basic and well-established approaches that are less computationally demanding, such as general dictionary and specific dictionary techniques.

This is particularly true for the literature surrounding earnings calls. The analysis conducted on earnings calls has demonstrated the importance of developing improved sentiment measures to further understand market movements, with considerable associations between earnings call sentiment and trading activity suggesting these exchanges between managers and analysts to be information dense. Though such associations are commonly statistically significant, the economic significance is comparatively weaker. A potential reason for the lack of economic significant findings could be because of a considerable proportion of the literature defining earnings calls sentiment using specific dictionary approaches. Furthermore, such studies typically focus solely on the text modality to define sentiment. The research of Mayew and Venkatachalam (2012) is something of an exception to this and suitably demonstrates leveraging the vocal modality on earnings calls to be informative.

Two main streams of future research stem from our synthesis of existing studies. Firstly, we expect to see the adoption and inclusion of state-of-the-art NLP techniques within academic finance over time, particularly in regard to the adoption of transformer architectures to classify the textual modality of financial text. Since the introduction of the transformer architecture by Vaswani et al. (2017), the adoption of the model has quickly solidified itself as the dominant architecture for NLP tasks (Wolf et al., 2020) with models such as BERT, GPT3 and T5 all adopting said architecture and returning state of the art results in a plethora of tasks (Nogueira & Cho, 2020; Sun et al., 2019). General pre-training of these models has been shown to produce strong performance in specific downstream tasks (Devlin et al., 2019; Howard & Ruder, 2018).

However, in a similar light to previous findings that specific word lists improve the understanding of sentiment for a specific context (e.g. Loughran & McDonald, 2011), transformer architecture becomes even more impressive with context-specific pre-training. Transformer models such as that used by Huang et al. (2023) are found to outperform other traditionally used methods and may be useful in generating higher classification accuracy for finance-specific contexts. Hence, it is necessary to implement the techniques that are achieving state-of-the-art results within the NLP literature to understand whether these measures can return more robust and economically significant relationships with market variables.

Secondly, leveraging both the text and audio modality on earnings calls to assess market characteristics has yet to be fully explored, yet offers exciting potential. The inclusion of nonverbal cues in finance academic literature is virtually absent (Mayew & Venkatachalam, 2012), whereas this modality has been extensively examined in other academic domains. For example, several studies within the psychology domain highlight the significance of vocal attributes influences on revealing the true underlying meaning of a message in the communication process. Mehrabian (1968) infers that 7% of human emotion is communicated through the semantic contents of a message, 38% through a message's vocal attributes and 55% via facial expression. This rule accentuates the lack of information conveyed through the textual modality alone.

In line with similar conclusions from psychology literature surrounding the importance of paralinguistic cues in the communication process, there are a number of studies employing sentiment analysis techniques that suggest a combination of text and audio data may improve classification accuracy, and consequently create a more robust representation of sentiment (Bhaskar et al., 2014; Dair et al., 2021; Houjeij et al., 2012; Yang et al., 2020). Hence, given that prior literature suggests both textual and vocal characteristics of earnings calls to be informative, and that Natural Language Processing literature finds a combination of text and audio to significantly increase classification accuracy, the adhesion of both measures represents a natural future direction for the literature.

This review, in accordance with the bulk of prior literature, has mainly focussed on the impact of sentiment on stock returns (Fisher et al., 2016). However, the conclusions drawn and future directions identified in this paper can also be applied to various other subdomains that leverage financial sentiment. For example, research on financial fraud detection (Goel & Gangolly, 2012; Goel & Uzuner, 2016; Humpherys et al., 2011; Moffitt & Burns, 2009) follows the same pattern as studies discussed within this review, in that most papers use dictionary and machine learning-based content analysis methods. This pattern continues throughout many other subdomains such as sentiment classification in different languages (Ghahfarrokhi & Shamsfard, 2020), identifying tax rates (Allen et al., 2021), assisting government reporting (Duan et al., 2022), defining blog sentiment (O'Leary, 2011) and crowdsourcing (O'Leary, 2016). The application of transformer architecture would likely benefit all these areas of research along with the leveraging of multiple modalities where data allow.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Andrew Todd https://orcid.org/0000-0001-7440-2342
James Bowden https://orcid.org/0000-0002-0419-1882

ENDNOTES

[1] A multimodal sentiment classifier is a model that leverages more than one modality of communication. The three modalities of communication are text, audio and visual.

[2] Also referred to as dictionaries or lexicons.

[3] https://rdrr.io/cran/SentimentAnalysis/man/DictionaryGI.html

[4] He uses the Wall Street Journal's (WSJ) frequently published 'Abreast of the Market' opinion piece as the source of news.

[5] Calculated using the GI on news articles, this pessimism factor is a linear combination of four categories from said dictionary namely: Negative, Weak, Fail and Fall word categories.

[6] DICTION is a dictionary-based language analysis programme that analyses the implied meaning of a text by searching it with the assistance of some 40 dictionaries or word lists (Given, 2008). The textual analysis software uses a series of five main dictionaries to search for sentiment features—Activity, Optimism, Certainty, Realism and Commonality—as well as 35 sub-features.

[7] Also see Stice (1991); Louis et al. (2008); Levi (2008).

[8] GPOMS is a textual content measure that quantifies mood in terms of six dimensions—Calm, Alert, Sure, Vital, Kind and Happy. OF is a mood tracking tool that measures positive vs negative mood. Both textual content measures fall under the category of general dictionaries.

[9] Facebook's Gross National Happiness (FGNH) indexes the positive and negative words used in the millions of status updates submitted daily by Facebook users. FGNH has face validity: it shows a weekly cycle and increases on national holidays. (Wang et al., 2012).

[10] See Siganos et al. (2014) for a list of these international markets.

[11] In their follow-up study (Siganos et al., 2017), the authors show a divergence of sentiment (DoS) measure returns strongly significant relationships between a contemporaneous daily increase in trading volume (2.829) and volatility (0.004)—a result which falls in line with Hirshleifer (1977), Harris and Raviv (1993) and Tetlock's (2007) findings that disagreement (in this case, portrayed through diverging sentiment) leads to increased trading because of market participants assigning different values to an asset.

[12] Sentiment accuracy suffers when general purpose dictionaries are used for specific domains that are not well represented by general language.

[13] Where financial results include unexpected earnings, a log of the market value of the firm's common equity, an indicator variable equal to one if earnings exceed analysts' forecast, and an indicator variable equal to one if earnings are greater than zero.

[14] Henry (2008) implies that past a certain point of positive tone, market reactions stop increasing. However, this specific level of tone is not defined.

[15] LM highlight that within financial disclosures negative words are more descriptive than positive words. This is because managers often convey negative news in positive words 'did not benefit' (p.38), whereas negative words rarely convey positive news.

[16] The monthly manager sentiment index is created by aggregating manager sentiment from 10Ks, 10Qs and conference call transcripts from 2003 to 2014.

[17] Further comparing their manager sentiment index to investor sentiment indexes (see p.131 for a list), the authors find that manager sentiment does not lead investor sentiment and vice versa. These findings indicate that manager sentiment and investor sentiment capture different subsets of sentiment information, and they are complementary in measuring market sentiment. Thus, manager sentiment has strong negative forecasting power for stock market returns. Jiang et al. (2019) conclude that the predictability found holds both in and out of sample showing its potential to generate economic value for investors.

[18] Expansion and Recession periods are taken from the National Bureau of Economic Research (NBER) Business Cycle Dating Committee. The NBER defines a recession as the period between a peak of economic activity and its subsequent trough. An expansion is defined between trough and peak.

[19] Comparison of the two periods suggests that expansionary periods are statistically different and return large economic differences—roughly three to four times stronger.

[20] The authors use four UK news sources: The Financial Times, The Times, The Guardian and The Mirror.

[21] Johnman et al. (2018) create a short-term reversal portfolio strategy. They take a long position on a stock if the previous days negative sentiment value is greater than the 70th percentile of last year's average negative sentiment value. Excluding transaction costs from 2002 to 2016 the sentiment strategy yields a greater return (0.061%) and Sharpe ratio (0.330) than a basic buy and hold strategy return (−0.007) and Sharpe ratio (−0.034). Ferguson et al. (2015) create a news-based trading strategy using positive and negative measures of sentiment as buy and sell signals. They create a long portfolio consisting of firms from the FTSE100 that have average net positive sentiment and a short portfolio comprised of firms that have net negative sentiment. Over the period 2003–2010 they returned 1.2 basis points per day, resulting in a significant alpha when all transaction costs were ignored.

[22] Evalauted over the three days surrounding the AGM (t-1 to t + 1),

[23] Evidenced through stronger market reactions from specific dictionary sentiment in comparison to general dictionary sentiment.

[24] The independence assumption states that features are independent of each other given the class. In an NLP setting, the naïve model assumes words are independent from each other.

[25] Antweiler and Frank (2004) classify messages as bullish, bearish or neither within this study to train their Naïve Bayes classifier. Bullish classifications correspond to messages that convey upward trends in prices or general metrics for a specific firm or an overall index. Bearish classifications are the opposite of bullish, and neutral classifications are messages that do not convey any significant information in regard to firm or market metrics.

[26] RNSE used NLP and ML techniques to produce a numerical indicator that classifies the relevance of news stories and news story sentiment.

[27] Blume et al. (1994) argue that higher volumes of media reflect a higher quality of news signal.

[28] News clustering relates to the production of news stories. The effect begins with an alert about the news content, and subsequent updates ultimately culminate in a full-blown story.

[29] Over periods of economic instability, the impact of these spill-over effects is intensified, which is broadly supportive of the earlier findings of Garcia (2012) who show that periods of heightened anxiety make investors more receptive to advice.

[30] Furthermore, considering contemporaneous movement factors and the VIX, tweet sentiment is still correlated with returns on the next day.

[31] Stock returns, trading volume, volatility, market capitalisation and the bid ask spread.

[32] El-Haj et al. (2019) cite that a potential reason for the lack of use of advanced NLP financial sentiment analysis models is the lack of substantial domain relevant datasets for training and testing.

[33] For a more detailed overview of transformer architecture, also see 'The annotated Transformer' by Vaswani et al. (2017), available at http://nlp.seas.harvard.edu/annotated-transformer/.

[34] The SQuAD Q&A task presents a model a paragraph with a question about the paragraph. The goal of the model is to effectively answer the question posed. The answers to the questions give insight to how well a model can understand text.

[35] ChatGPT is built upon a variant of this model.

[36] The LAMBADA dataset tests a model's ability to handle long-range dependencies in text. The task requires a model to predict the last word of a sentence based upon a context paragraph as input.

[37] The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training and evaluating NLP models.

[38] They compare with a RNN based Bidirectional LSTM, BiLSTM (see Hochreiter & Schmidhuber, 1997), the Multichannel Convolutional Neural Network (CNN) (see Kim, 2014), the CPU-efficient FastText (see Joulin et al., 2016) that is adopted by Facebook, and the Transformer with attention mechanism (see Vaswani et al., 2017).

[39] Precision score is the number of positive class predictions that belong to the positive class. Recall score is the number of positive class predictions from all positive examples in the dataset. F1 score is a measure of balance that concerns both precision and recall in one number.

[40] A publicly available financial sentiment dataset consisting of 4840 sentences from financial news.

[41] A dataset consisting of 10,000 sentences labelled with positive, negative or neutral sentiment taken from analyst reports.

[42] An open challenge dataset consisting of 1111 sentences annotated for financial sentiment.

[43] However, the authors do note that this branch of sentiment analysis, although promising, is still in its infancy.

[44] Poria et al. (2015) further evaluate the accuracy of combinations of two modalities against the trimodal model and each modality on its own. The authors find that all dual combinations of data outperform all singular modality models. They highlight that, absent the trimodal model, a combination of visual and audio data performs the best then textual and visual data next best and finally textual and audio data performing the worst out of all pairs of modalities but still better than any singular modality model.

[45] Guyer et al. (2018) note a great deal of research has revealed that the content of what we say matters, indicating that the way in which we communicate matters. More succinctly, how we speak conveys substantial information beyond the content of communication. There is a substantial body of psychology literature that relates to vocal characteristics and their impact on persuasion/decision-making. For example, prior literature has shown that vocal pitch impacts listeners' perception of speakers' personal traits and qualities. Qualities such as credibility, tranquility, persuasion, trustworthiness and maturity are associated with a lower level of vocal pitch. Conversely, high pitch voices are considered immature, nervous, informal, less credible and less persuasive (Chattopadhyay et al., 2003; Chua et al., 2020; Feinberg et al., 2005; Klofstad et al., 2012; Martín-Santana et al., 2015; Song et al., 2020; Wang et al., 2018). Research has also been conducted on vocal intonation (Apple et al., 1979; Brooke & Ng, 1986; Gélinas-Chebat et al., 1996; Wallbott, 1982), intensity (Bradac et al., 1988; Brooke & Ng, 1986; Conley et al., 1978; Erickson et al., 1978), jitter and shimmer (Giddens et al., 2013; Mendoza & Carballo, 1998; Park et al., 2011) showing how variations of these features impact speaker persuasion and listener perceptions/decision-making.

[46] Paralinguistic data can be generated from earnings conference calls using speech analysis software. For instance, in the research conducted following this paper, paralinguistic data were created for the S&P100 by forcefully aligning earnings calls transcripts and their corresponding audio files to get sentence level audio clips and then applied to PRAAT to generate paralinguistic data. The paralinguistic data created, accompanied by numerical representations of the text generated using a Transformer, were then fed to a neural network to classify sentiment.

[47] The first component involves the discussion of firm performance by executives, and the second focusses on a question-and-answer session between firm executives and market participants.

[48] The three-day window from the day before to the day after an earnings call.

[49] From 2 days after a call up to 60 days after a call.

[50] The authors note that REITs are constantly involved in asset acquisition and disposition activities. Hence, the underlying revenue generating

asset bases are constantly changing for REITs. These unique characteristics of REITs provide a natural setting in which to study the relationship between stock returns and conference call content.

[51] Calculated using the GI software – general dictionary approach.

[52] Low earnings surprise is when a firms' reported profits are significantly below its earnings estimate.

[53] The authors evaluate an options market instead of the commonly assessed equities markets. A share price in an equities market reflects the current value of the firm. However, the implied volatilities in an options market reflect investors' uncertainty surrounding a firms' future value. Hence, Borochin et al. (2017) evaluate value uncertainty.

[54] The pessimism and optimism factors are calculated using a modified LM dictionary.

[55] A one standard deviation movement in managerial introductory tone reflects a −0.010 movement in the value uncertainty measure.

[56] A one standard deviation movement in analyst Q&A tone incites a −0.028 movement in the value uncertainty measure.

[57] Calculated using the LM dictionary approach.

[58] Extreme downside risk in returns. It can be defined as the conditional skewness of the returns distribution which captures the information asymmetry, between inside managers and outside investors, in the risk associated with a stock.

[59] Regression results for the predictive power of call tone on stock price crash risk indicate that an increase of one standard deviation in call tone results in a decrease in stock price crash risk of 0.092 standard deviations.

[60] The underlying emotional state. An example of positive affective states is defined by the author as happiness, excitement and enjoyment. Examples of negative affective states are fear, tension and anxiety. These states are defined using Layered Voice Analysis software.

[61] The length of a period is 90 days.

[62] Managers' sentiment as defined by the Henry (2006) and Loughran and McDonald (2011) word lists enhance prediction of future operating performance.

[63] Manager specific sentiment can be thought of like a personality. Everyone has a personality specific to them and it does not change no matter the job or situation they are in. Davis et al. (2015) evaluate whether managers have a specific sentiment, much like their personality, and whether this can be identified throughout different roles in their career.

[64] Loughran and McDonald's (2011) word list.

[65] Both defined using LM positive and negative word lists.

[66] Based on these findings, the researchers attempt to create a portfolio based upon analyst sentiment measures. They create the portfolio by 'going long' on stocks with high analyst sentiment and 'going short' on stocks with low analyst sentiment. This produced a significant abnormal return of 1.32% over a 6-month period. Thus, the authors conclude that a portfolio based upon this strategy is a good idea and, however, may not be economically significant considering the inclusion of transaction costs.

[67] Defined using LM positive and negative word lists.

[68] This effect is not seen with managerial sentiment.

[69] This does not mean that managerial sentiment should be disregarded however as it is still influential.

[70] This measure is the difference between introductory statement sentiment and Q&A sentiment. The authors note that it measures inflated talk by managers who often speak overly optimistically in the introductory statements in comparison to more objective talk from analysts in the Q&A section. Kartik et al. (2007) identify that inflated talk should be considered bad news—hence, the authors are evaluating whether short sellers are sophisticated enough to process inflated talk information in their forecasts as bad news.

[71] Bochkay et al. (2020) create a corpus of extreme words by deploying a Human Intelligence Task (HIT) on Amazon's Mechanical Turk service (MTurk). This task asked 'highly qualified English-speaking workers' to rank 50 randomly selected words from the author's dictionary on a scale from −5 (*extremely negative*) to +5 (*extremely positive*). The average score from all participants was used to rank extreme words.

REFERENCES

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

Alamoudi, E., & 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), 259–281. https://doi.org/10.1080/12460125.2020.1864106

Allen, E., O'Leary, D. E., Qu, H., & Swenson, C. W. (2021). Tax specific versus generic accounting-based textual analysis and the relationship with effective tax rates: Building context. *Journal of Information Systems*, *35*(2), 115–147. https://doi.org/10.2308/ISYS-2020-018

Amoozegar, A., Berger, D., Cao, X., & Pukthuanthong, K. (2020). Earnings conference calls and institutional monitoring: Evidence from textual analysis. *Journal of Financial Research*, *43*(1), 5–36. https://doi.org/10.1111/jfir.12199

Antweiler, W., & Frank, M. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, *37*(5), 715–727. https://doi.org/10.1037/0022-3514.37.5.715

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 492–499). IEEE.

Audrino, F., & Tetereva, A. (2019). Sentiment spillover effects for US and European companies. *Journal of Banking & Finance*, *106*, 542–567. https://doi.org/10.1016/j.jbankfin.2019.07.022

Azar, P., & Lo, A. (2016). The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds. *The Journal of Portfolio Management*, *42*(5), 123–134. https://doi.org/10.3905/jpm.2016.42.5.123

Bannier, C., Pauls, T., & Walter, A. (2017). *CEO-speeches and stock returns* (Vol. 583). Center for Financial Studies.

Bernanke, B., & Kuttner, K. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of Finance*, *60*(3), 1221–1257. https://doi.org/10.1111/j.1540-6261.2005.00760.x

Bhaskar, J., Sruthi, K., & Nedungadi, P. (2014). Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. In *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)* (pp. 1–6). IEEE.

Bhonde, R., Bhagwat, B., Ingulkar, S., & Pande, A. (2015). Sentiment analysis based on dictionary approach. *International Journal of Emerging Engineering Research and Technology*, *3*(1), 51–54.

Blau, B., DeLisle, J., & Price, S. (2015). Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. *Journal of Corporate Finance*, *31*, 203–219. https://doi.org/10.1016/j.jcorpfin.2015.02.003

Blume, L., Easley, D., & O'Hara, M. (1994). Market statistics and technical analysis: The role of volume. *The Journal of Finance*, *49*(1), 153–181. https://doi.org/10.1111/j.1540-6261.1994.tb04424.x

Bochkay, K., Hales, J., & Chava, S. (2020). Hyperbole or reality? Investor response to extreme language in earnings conference calls. *The Accounting Review*, *95*(2), 31–60. https://doi.org/10.2308/accr-52507

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Borochin, P., Cicon, J., DeLisle, R., & Price, S. (2017). The effects of conference call tones on market perceptions of value uncertainty. *Journal of Financial Markets*, *40*, 75–91. https://doi.org/10.1016/j.finmar.2017.12.003

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 223–232).

Bowden, J., Kwiatkowski, A., & Rambaccussing, D. (2019). Economy through a lens: Distortions of policy coverage in UK national newspapers. *Journal of Comparative Economics*, *47*(4), 881–906. https://doi.org/10.1016/j.jce.2019.07.002

Bowen, R. M., Davis, A. K., & Matsumoto, D. A. (2002). Do conference calls affect analysts' forecasts? *The Accounting Review*, *77*(2), 285–316. https://doi.org/10.2308/accr.2002.77.2.285

Bradac, J. J., Mulac, A., & House, A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluative consequences. *Language and Communication*, *8*(3–4), 213–228. https://doi.org/10.1016/0271-5309(88)90019-5

Brockman, P., Li, X., & Price, S. (2015). Differences in conference call tones: Managers vs. analysts. *Financial Analysts Journal*, *71*(4), 24–42. https://doi.org/10.2469/faj.v71.n4.1

Brooke, M. E., & Ng, S. H. (1986). Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, *5*(3), 201–210. https://doi.org/10.1177/0261927X8600500303

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research [Review Article]. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227

Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27. https://doi.org/10.5117/CCR2021.1.001.CHAN

Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, *94*, 53–64. https://doi.org/10.1016/j.dss.2016.10.006

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, *70*(2), 223–260. https://doi.org/10.1016/S0304-405X(03)00146-6

Chattopadhyay, A., Dahl, D. W., Ritchie, R. J., & Shahin, K. N. (2003). Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology*, *13*(3), 198–204. https://doi.org/10.1207/S15327663JCP1303_02

Chen, J., Nagar, V., & Schoenfeld, J. (2018). Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, *23*(4), 1315–1354. https://doi.org/10.1007/s11142-018-9453-3

Chua, G. Y. P., Er, H. J., Liaw, S. Y., & He, T. S. (2020). Pitch right: The effect of vocal pitch on risk aversion. *Economics Bulletin*, *40*(4), 3131–3139.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modelling. In *Deep learning and representation learning workshop*. Neural Information Processing Systems.

Cieslak, A., Morse, A., & Vissing-Jorgensen, A. (2014). *Stock returns over the FOMC cycle*. NBER Working Paper. https://doi.org/10.2139/ssrn.2687614

Conley, J. M., O'Barr, W. M., & Lind, E. A. (1978). *The power of language: Presentational style in the courtroom* (Vol. 1978) (p. 1375). Duke Lj. https://doi.org/10.2307/1372218

Dair, Z., Donovan, R., & O'Reilly, R. (2021). Classification of emotive expression using verbal and non-verbal components of speech. In *2021 32nd Irish Signals and Systems Conference (ISSC)* (pp. 1–8). IEEE.

D'Andrea, E., Ducange, P., Bechini, A., Renda, A., & Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, *116*, 209–226. https://doi.org/10.1016/j.eswa.2018.09.009

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389. https://doi.org/10.1016/j.knosys.2021.107389

Davis, A., Ge, W., Matsumoto, D., & Zhang, J. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, *20*(2), 639–673. https://doi.org/10.1007/s11142-014-9309-4

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2012). The effect of managerial "style" on the tone of earnings conference calls. In *CAAA Annual Conference*. Retrieved from http://www.usc.edu/schools/business/FBE/seminars/papers/ARF_9-21-12_GE.pdf

Davis, A., & Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A*. *Contemporary Accounting Research*, *29*(3), 804–837. https://doi.org/10.1111/j.1911-3846.2011.01125.x

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. *arXiv preprint arXiv:1810.04805*.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using Naïve Bayes' and K-NN classifier. *International Journal of Information Engineering and Electronic Business*, *8*(4), 54–62. https://doi.org/10.5815/ijieeb.2016.04.07

Diesner, J., & Evans, C. (2015). Little bad concerns: Using sentiment analysis to assess structural balance in communication networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 342–348).

Doran, J., Peterson, D., & Price, S. (2012). Earnings conference call content and stock price: The case of REITs. *The Journal of Real Estate Finance and Economics*, *45*(2), 402–434. https://doi.org/10.1007/s11146-010-9266-z

Duan, H. K., Hu, H., Yoon, Y., & Vasarhelyi, M. (2022). Increasing the utility of performance audit reports: Using textual analytics tools to improve government reporting. *Intelligent Systems in Accounting, Finance and Management*, *29*(4), 201–218. https://doi.org/10.1002/isaf.1526

El-Haj, M., Rayson, P., Walker, M., Young, S., & 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, E. A., Johnson, B. C., & O'Barr, W. M. (1978). Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of Experimental Social Psychology*, *14*(3), 266–279. https://doi.org/10.1016/0022-1031(78)90015-X

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, *69*(3), 561–568. https://doi.org/10.1016/j.anbehav.2004.06.012

Ferguson, N., Philip, D., Lam, H., & Guo, J. (2015). Media content and stock returns: The predictive power of press. *Multinational Finance Journal*, *19*(1), 1–31. https://doi.org/10.17578/19-1-1

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214. https://doi.org/10.1002/isaf.1386

Frankel, R., Johnson, M., & Skinner, D. J. (1999). An empirical examination of conference calls as a voluntary disclosure medium. *Journal of*

Accounting Research, *37*(1), 133–150. https://doi.org/10.2307/2491400

Fu, X., Wu, X., & Zhang, Z. (2019). The information role of earnings conference call tone: Evidence from stock price crash risk. *Journal of Business Ethics*, *173*, 643–660. https://doi.org/10.1007/s10551-019-04326-1

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444. https://doi.org/10.1016/j.inffus.2022.09.025

Garcia, D. (2012). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267–1300.

Gélinas-Chebat, C., Chebat, J. C., & 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), 243–262. https://doi.org/10.2466/pms.1996.83.1.243

Ghahfarrokhi, A., & Shamsfard, M. (2020). Tehran stock exchange prediction using sentiment analysis of online textual opinions. *Intelligent Systems in Accounting, Finance and Management*, *27*(1), 22–37. https://doi.org/10.1002/isaf.1465

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: a review. *Journal of Voice*, *27*(3), 390–e21.

Given, L. (2008). *The SAGE encyclopedia of qualitative research methods*. DICTION (Software). https://doi.org/10.4135/9781412963909

Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, *19*(2), 75–89. https://doi.org/10.1002/isaf.1326

Goel, S., & Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 215–239. https://doi.org/10.1002/isaf.1392

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication. *The Annals of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In *Information and communication technologies in tourism 2012* (pp. 460–470). Springer.

Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, *18*(2), 321–340. https://doi.org/10.1016/j.jempfin.2010.11.009

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, *121*, 105969. https://doi.org/10.1016/j.jbankfin.2020.105969

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, *2*(3), 153–170. https://doi.org/10.1016/j.jfds.2017.02.001

Guyer, J. J., Fabrigar, L. R., Vaughan-Johnston, T. I., & Tang, C. (2018). The counterintuitive influence of vocal affect on the efficacy of affectively-based persuasive messages. *Journal of Experimental Social Psychology*, *74*, 161–173.

Harris, M., & Raviv, A. (1993). Differences of opinion make a horse race. *Review of Financial Studies*, *6*(3), 473–506. https://doi.org/10.1093/rfs/5.3.473

Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting.*, *3*(1), 1–19. https://doi.org/10.2308/jeta.2006.3.1.1

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, *45*(4), 363–407. https://doi.org/10.1177/0021943608319388

Hiew, J., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). *BERT-based financial sentiment index and LSTM-based stock return predictability*. Cornell University Working Paper.

Hirshleifer, J. (1977). Economics from a biological viewpoint. *The Journal of Law and Economics*, *20*(1), 1–52. https://doi.org/10.1086/466891

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Houjeij, A., Hamieh, L., Mehdi, N., & Hajj, H. (2012). A novel approach for emotion classification based on fusion of text and speech. In *2012 19th International Conference on Telecommunications (ICT)* (pp. 1–6). IEEE.

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146*.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009

Jegadeesh, N., & Wu, A. (2012). Word power: A new approach for content analysis. *Journal of Financial Economics*, *3*(110), 712–729. https://doi.org/10.1016/j.jfineco.2013.08.018

Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, *132*(1), 126–149. https://doi.org/10.1016/j.jfineco.2018.10.001

Jiang, W. (2021). Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, *184*, 115537. https://doi.org/10.1016/j.eswa.2021.115537

Johnman, M., Vanstone, B., & Gepp, A. (2018). Predicting FTSE 100 returns and volatility using sentiment analysis. *Accounting and Finance*, *58*(S1), 253–274. https://doi.org/10.1111/acfi.12373

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651*.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410*.

Kartik, N., Ottaviani, M., & Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic Theory*, *134*(1), 93–116. https://doi.org/10.1016/j.jet.2006.04.003

Kaushik, L., Sangwan, A., & Hansen, J. H. (2013). Sentiment extraction from natural audio streams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8485–8489). IEEE.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185. https://doi.org/10.1016/j.irfa.2014.02.006

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Klofstad, C. A., Anderson, R. C., & 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), 2698–2704. https://doi.org/10.1098/rspb.2012.0311

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99–117. https://doi.org/10.1007/s10772-011-9125-1

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

Levi, S. (2008). Voluntary disclosure of accruals in earnings press releases and the pricing of accruals. *Review of Accounting Studies*, *13*(1), 1–21. https://doi.org/10.1007/s11142-007-9059-7

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), 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Louis, H., Robinson, D., & Sbaraglia, A. (2008). An integrated analysis of the association between accrual disclosure and the abnormal accrual anomaly. *Review of Accounting Studies*, *13*(1), 23–54. https://doi.org/10.1007/s11142-007-9038-z

Lucca, D., & Moench, E. (2015). The pre-FOMC announcement drift. *The Journal of Finance*, *70*(1), 329–371. https://doi.org/10.1111/jofi.12196

Luong, M. T., Pham, H., & Manning, C. D. (2015). *Effective approaches to attention-based neural machine translation*. *arXiv preprint arXiv:1508.04025*.

Mairesse, F., Polifroni, J., & Di Fabbrizio, G. (2012). Can prosody inform sentiment analysis? experiments on short spoken reviews. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5093–5096). IEEE.

Mao, H., Counts, S., & Bollen, J. (2011). *Predicting financial markets: Comparing survey, news, twitter and search engine data*. *arXiv preprint arXiv:1112.1051*.

Martín-Santana, J. D., Muela-Molina, C., Reinares-Lara, E., & Rodríguez-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), 143–160. https://doi.org/10.1016/j.brq.2014.06.001

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

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

McGurk, Z., Nowak, A., & Hall, J. (2020). Stock returns and investor sentiment: textual analysis and social media. *Journal of Economics and Finance*, *44*(3), 458–485. https://doi.org/10.1007/s12197-019-09494-4

McKay Price, S., Doran, J., Peterson, D., & Bliss, B. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992–1011. https://doi.org/10.1016/j.jbankfin.2011.10.013

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mehrabian, A. (1968). Inference of attitudes from the posture, orientation, and distance of a communicator. *Journal of Consulting and Clinical Psychology*, *32*(3), 296–308. https://doi.org/10.1037/h0025906

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, *12*(3), 263–273. https://doi.org/10.1016/S0892-1997(98)80017-9

Milian, J., & Smith, A. (2017). An investigation of analysts' praise of management during earnings conference calls. *Journal of Behavioral Finance*, *18*(1), 65–77. https://doi.org/10.1080/15427560.2017.1276068

Moffitt, K., & Burns, M. B. (2009). What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. In *AMCIS 2009 Proceedings* (p. 399).

Morency, L.-P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169–176).

Munikar, M., Shakya, S., & Shrestha, A. (2019). Fine-grained sentiment classification using BERT. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)* (Vol. 1, pp. 1–5). IEEE.

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys*, *30*(2), 356–369. https://doi.org/10.1111/joes.12102

Nogueira, R., & Cho, K. (2020). *Passage re-ranking with BERT*. Cornell University Working Paper.

O'Leary, D. E. (2011). Blog mining-review and extensions: "From each according to his opinion". *Decision Support Systems*, *51*(4), 821–830. https://doi.org/10.1016/j.dss.2011.01.016

O'Leary, D. E. (2016). On the relationship between number of votes and sentiment in crowdsourcing ideas and comments for innovation: A case study of Canada's digital compass. *Decision Support Systems*, *88*, 28–37. https://doi.org/10.1016/j.dss.2016.05.006

Park, C. K., Lee, S., Park, H. J., Baik, Y. S., Park, Y. B., & Park, Y. J. (2011). Autonomic function, voice, and mood states. *Clinical Autonomic Research*, *21*, 103–110. https://doi.org/10.1007/s10286-010-0095-1

Pereira, J., Luque, J., & Anguera, X. (2014). Sentiment retrieval on web reviews using spontaneous natural speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4583–4587). IEEE.

Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2539–2544).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25–40. https://doi.org/10.1016/j.jbankfin.2017.07.002

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), 1–13. https://doi.org/10.1007/s42521-019-00014-x

Ribeiro, F., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization*, *107*, 730–743. https://doi.org/10.1016/j.jebo.2014.06.004

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2017). Divergence of sentiment and stock market trading. *Journal of Banking & Finance*, *78*, 130–141. https://doi.org/10.1016/j.jbankfin.2017.02.005

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, *65*, 3–14. https://doi.org/10.1016/j.imavis.2017.08.003

Song, S., Baba, J., Nakanishi, J., Yoshikawa, Y., & Ishiguro, H. (2020). Mind the voice!: Effect of robot voice pitch, robot voice gender, and user gender on user perception of teleoperated robots. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–8).

Sprenger, T., Tumasjan, A., Sandner, P., & Welpe, I. (2013). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926–957. https://doi.org/10.1111/j.1468-036X.2013.12007.x

Stice, E. (1991). The market reaction to 10-K and 10-Q filings and to subsequent The Wall Street journal earnings announcements. *The Accounting Review*, *66*(1), 42–55.

Stone, P., & Hunt, E. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference* (pp. 241–256).

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming*, China, October 18–20, 2019, Proceedings 18 (pp. 194–206). Springer International Publishing.

Sun, L., Najand, M., & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, *73*, 147–164. https://doi.org/10.1016/j.jbankfin.2016.09.010

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Troussas, C., Virvou, M., Espinosa, K., Llaguno, K., & Caro, J. (2013). *Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning*. In IISA 2013 (pp. 1–6). IEEE.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media* (Vol. 4, No. 1, pp. 178–185).

Twedt, B., & Rees, L. (2012). Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy*, *31*(1), 1–21. https://doi.org/10.1016/j.jaccpubpol.2011.10.010

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. Advances in neural information processing systems, 30.

Wallbott, H. G. (1982). Contributions of the German "expression psychology" to nonverbal communication research: Part III: Gait, gestures, and body movement. *Journal of Nonverbal Behavior*, *7*, 20–32. https://doi.org/10.1007/BF01001775

Wang, N., Kosinski, M., Stillwell, D., & Rust, J. (2012). Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research*, *115*(1), 483–491. https://doi.org/10.1007/s11205-012-9996-9

Wang, T. Y., Kawaguchi, I., Kuzuoka, H., & Otsuki, M. (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*, *2*(CSCW), 1–18.

Wang, W., & Hua, Z. (2014). A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Volume 1: Long Papers, pp. 1155–1165).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45).

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., & Klingner, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144*.

Yang, K., Xu, H., & Gao, K. (2020). CM-BERT: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 521–528).

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