Measuring Financial Tone in Earnings Calls

An evaluation of FinBERT

Luuk Hopman

Student number: u1273783

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Data Science and Society
Department of Cognitive Science and Artificial Intelligence
School of Humanities and Digital Sciences

Thesis committee

Supervisor: Dr. G. Chrupala

Second reader: Dr. R.G. Alhama

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands

May 2021

Contents

Measuring Financial Tone in Earnings Calls

An evaluation of FinBERT

Luuk Hopman
l.a.m.hopman@tilburguniversity.edu
Tilburg University
Tilburg, The Netherlands

Abstract

Prior research in the financial text sentiment literature measures tone using positive and negative word lists from sources such as the Loughran & McDonald Dictionary. This approach fails to capture the deeper semantics of texts. In this paper, I evaluate the capabilities of the recently introduced FinBERT language model as an alternative tone classifier. First, I show that FinBERT significantly outperforms sentiment lexicons on two labelled data sets. The superior performance is attributable to the fact that, unlike dictionaries, FinBERT considers linguistic context. Second, using a sample of over 25,000 earnings calls, I analyse the statistical power and economic impact of the different sentiment algorithms with market reaction and analysts output models. The results indicate that the finance-specific sentiment dictionary underestimates the economic magnitude of textual tone compared to FinBERT. Additionally, I demonstrate that the measurement error of lexicon-based approaches is especially problematic with smaller sample sizes. Last, I document novel evidence for managerial tone manipulation in corporate disclosures.

Keywords Natural Language Processing · Textual Tone · Deep Learning · Corporate Disclosures · Capital Markets

1 Introduction

In the last decades, numerous studies in the accounting and finance literature have used linguistic tone to study corporate disclosure. They demonstrate the impact of tone, commonly defined as the ratio between positively and negativity in a text, in relation to financial variables such as stock returns forecasting (Price, Doran, Peterson, & Bliss, 2012; Schumaker & Chen, 2009), financial distress prediction (Hájek & Olej, 2013), and fraud detection (Purda & Skillicorn, 2015). These studies show the value of text sentiment analysis in the financial domain. However, despite the advances in natural language processing (NLP) sentiment models, accounting and finance researchers rarely adopt modern sentiment models in studies.

Sentiment lexicons, or dictionaries, are the most prominent method for sentiment classification in financial text research (e.g. Cao, Jiang, Yang, & Zhang, 2020; Chen, Nagar, & Schoenfeld, 2018; Henry & Leone, 2016). Generally, a sentiment dictionary consists of word collections that share a sentiment orientation such as positivity, negativity, and uncertainty. This approach ignores the complex nature of languages by treating words independently without comprehending the context. The word *default* has a substantially different meaning in "debt default" compared to "default choice". Bag of words machine learning models that are sometimes employed (e.g. Li, 2010) share this drawback, as they deal with words as independent tokens. Overall, the sentiment algorithms that the vast majority of studies use fail to capture deeper lexical semantics.

Another potential drawback of sentiment word lists is that they allow for easy manipulation. Anecdotal evidence points out a so-called NLP arms race between managers and outsiders: "We are not far from someone on a call reading *we said au revoir to our profitability* versus *we recorded a loss* because it reads better in some NLP model" (Wigglesworth, 2020). In agreement with this, a recent study finds that managers shift their tone as a response to the surge in robotic trading based on CEO sentiment (Cao et al., 2020). This suggests that word lists are getting less reliable for tone analysis in financial texts as corporate insiders learn to exploit them. More dynamic models are arguably more challenging to manipulate than static word lists.

NLP research has introduced neural networks that can overcome these deficiencies. A popular example is the Long Short-Term Memory (LSTM) model (Gers, Schmidhuber, & Cummins, 1999), a recurrent neural network that is designed to consider the context of a word based on words that surround it. A more recent NLP advancement is the Transformer (Vaswani et al., 2017). Like LSTMs, the Transformer is designed to handle sequences of data. However, the structure of Transformers allows for processing based on the entirety of the sequence, such as a sentence,

rather than word by word. Nonetheless, models that consider context are theoretically superior to those that do not. Unsurprisingly, therefore, computer scientists document that both LSTM and Transformer models outperform bag-of-words models on sentiment analysis exercises.

The complexity of deep learning NLP models can be considered an obstacle. A good understanding of programming languages like Python, that many researchers in social sciences lack, is needed to assimilate such a model. Additionally, most popular deep learning algorithms demand a large train data set for optimal performance. Moreover, deep neural network architectures that show good results on sentiment analysis tasks, such as a CNN-LSTM model (Wang, Yu, Lai, & Zhang, 2016), have millions of parameters. Therefore, training and applying neural networks requires considerable computing power. In consequence, some NLP researchers question whether the gain in accuracy achieved by state-of-the-art models outweighs the environmental impact of the needed computational resources (Strubell, Ganesh, & McCallum, 2019).

One of the potential solutions is to employ pre-trained neural networks for NLP tasks. In that way, many researchers and practitioners benefit from the same trained model. A well-known example is the Generative Pre-trained Transformer 3 (GPT-3) auto-regressive language model that was announced in June 2020 by OpenAI. Brown et al. (2020) document successful applications on GPT-3 in various NLP subfields including sentiment classification, text generation, and translation. However, at this moment GPT-3 lacks accessibility. An open-source alternative is Bidirectional Encoder Representations from Transformers (hereafter BERT) developed by Google Research (Devlin, Chang, Lee, & Toutanova, 2019). Previous literature shows that BERT achieves state-of-the-art performances on downstream NLP tasks (Zhu et al., 2019; van Aken, Winter, Löser, & Gers, 2019; Munikar, Shakya, & Shrestha, 2019).

However, there is no consensus on whether deep learning algorithms can outperform domain-specific sentiment dictionaries on financial text classification exercises. The extant research suggests that texts in the financial domain differ significantly from general texts. Financial texts have a unique vocabulary that requires a domain-specific NLP approach (Mishev, Gjorgjevikj, Vodenska, Chitkushev, & Trajanov, 2020). Researchers have introduced sentiment word lists for financial disclosures (Loughran & McDonald, 2011, hereafter "LM") and demonstrate that they outperform more complex sentiment classification models (Henry & Leone, 2016). Accordingly, it is crucial to measure a model's performance on financial texts specifically.

In line with that, an interesting recent trend in pre-trained NLP model research is domain-specific models. For example, BioBERT (Lee et al., 2020) is a variant of BERT designed for biomedical text mining. Similarly, Yang, UY, and Huang (2020) introduce FinBERT, a BERT-

based model for financial communications. The authors train it on a large corpus of financial texts extracted from annual reports (10-Ks), quarterly reports (10-Qs), earnings call transcripts, and analyst reports. FinBERT is publicly available to researchers and practitioners.

This paper evaluates the capabilities of FinBERT as a financial text sentiment classifier. In this paragraph, I formulate the research questions addressed by this study. First, I contrast the performance of two popular sentiment dictionaries in the finance and accounting literature, the Harvard IV-4 Dictionary and the LM Dictionary, to FinBERT on annotated financial phrases.

(RQ1) *To what extent can FinBERT improve the accurateness of financial text sentiment classification on labelled financial phrases compared to dictionary-based approaches?*

Second, this study aims to shed some light on whether FinBERT is a higher-quality substitute to dictionaries in financial textual tone research. Therefore, I thoroughly examine the statistical significance and relative predictive power (economic impact) of FinBERT compared to the LM Dictionary in multiple empirical sentiment classification settings using a panel data set with over 25,000 earnings conference calls of S&P 500 firms from 2004 to 2020. Specifically, I use the tone of the question and answer (Q&A) component of the call, measured by the LM lexicon and FinBERT, to test the predictive capabilities of the algorithms. The market reaction and analyst output (stock recommendation levels, price targets, and forecasts) following the call serve as proxies for the underlying sentiment of the calls.

(RQ2) *What is the difference in statistical power between FinBERT and the LM Dictionary as textual tone measures in empirical finance and accounting research?*

(RQ3) *What is the economic impact of using FinBERT as textual tone measure compared to the LM Dictionary?*

Last, following Cao et al. (2020), I analyse whether managers adjusted their vocabulary to the deceive NLP algorithms based on the LM Dictionary. It is crucial to understand this as it potentially results in biased tone measures. Thus, disclosure tone manipulation may serve as another reason to move avoid dictionary-based methods.

(RQ4) *Do corporate insiders avoid negative words in the LM Dictionary?*

The remainder of this paper is organised as follows. Section 2 reviews the relevant literature and specifies my contributions. Section 3 describes the data, sample, and model. Section 4 documents the conducted experiments. Section 5 presents the empirical results. Section 6 discusses and concludes the study.

2 Related Literature

2.1 Financial Communications

Companies issue both quantitative (e.g. balance sheets and income statements) and qualitative (e.g. Management Discussion and Analysis and conference calls) disclosures to their stakeholders. The extant literature suggests that qualitative information is necessary for investors to interpret quantitative information (Davis, Piger, & Sedor, 2012; Huang, Zang, & Zheng, 2014). Accordingly, there is a growing academic interest in textual financial communications.

Earnings calls are characterised as voluntary corporate disclosures and a primary source of management-analyst interactions. The majority of U.S. firms hold quarterly earnings conference calls to discuss the quarterly earnings report (Hollander, Pronk, & Roelofsen, 2010). Researchers argue that earnings calls are one of the media through which firms channel value-relevant information to investors (Kimbrough, 2005). Previous studies suggest that these calls are informative for stakeholders including auditors, buy-side analysts, and individual investors (e.g. Matsumoto, Pronk, & Roelofsen, 2011; Jung, Wong, & Zhang, 2018). Moreover, conference calls reduce the information asymmetry between the management and investors (Brown, Hillegeist, & Lo, 2004).

In a typical earnings call, the management first presents the firm's quarterly results and provides an outlook to the future. After this presentation, analysts ask questions during the Q&A section of the call. Matsumoto et al. (2011) find that both components of the call are informative for the market. However, the Q&A segment generally conveys more information than the management presentation, especially for firms with larger analysts followings. This is not surprising since the introductory presentation is usually a scripted reiteration of the earnings press release (Kimbrough, 2005). The meaningful role of analysts in earnings calls is further accentuated by the finding that active analysts' participation in conference calls increases the informativeness of the call (Matsumoto et al., 2011).

2.2 Sentiment Analysis in Finance

The use of positive and negative communication is a topic with a burgeoning interest in behavioural finance. Many studies document the economic impact of tone in financial texts. For example, the extant literature reports that a positive tone in corporate disclosures results in a stock price increase in the short term (Davis et al., 2012; Huang, Zang, & Zheng, 2014). This positive market reaction reverses in the long term (Huang, Zang, & Zheng, 2014). These studies document a significant impact of linguistic tone, indicating that relevant information is conveyed

through word choices in disclosures.

Due to its qualitative nature, it is challenging to quantify textual tone in a large sample. Lexicons are the most prominent approach of gauging linguistic tone in the finance and accounting literature. Typically, studies define tone as the percentage of positive and negative words in a text or as the difference between positive and negative words scaled by the number of tone words (e.g. Henry & Leone, 2016). Many studies (e.g. Henry & Leone, 2016; Huang, Teoh, & Zhang, 2014; Price et al., 2012) employ the positive and negative word lists of the general-purpose Harvard IV-4 Psychosocial lexicon. However, the use of general tone dictionaries in accounting and finance research receives criticism in the extant literature as they include words that are irrelevant in financial texts (Henry & Leone, 2016; Loughran & McDonald, 2011). Loughran and McDonald (2011) argue that the difference between financial and general language is problematic. They find that the majority of the words that are classified as negative in a general tone dictionary do not carry a negative connotation in the financial domain. For example, *beat* is a pessimistic word in general language, but often positive in the financial dialect (e.g. "Beat the consensus estimates"). On the other hand, financial terminology with a clear sentimental tone, such as *outperform* (*underperform*), are neutral words in general-purpose dictionaries. Consequently, Loughran and McDonald introduce a domain-specific sentiment lexicon. The LM Dictionary consists of word lists that consider the meaning of words in a financial context. Its positive and negative word lists are jointly the most prominent textual tone measure in accounting and finance literature (Luo & Zhou, 2020).

Machine learning algorithms are not prevalent in the financial sentiment literature (Luo & Zhou, 2020). A notable exception is Li (2010), who presents a Naive Bayes machine learning classifier as an alternative to word lists. In contrast with the popularity of deep learning NLP models like LSTMs and Transformers in other domains, a recent financial tone literature review does not report a single study that incorporates deep learning (Luo & Zhou, 2020). The lack of machine learning in financial tone research can be explained by multiple factors. First, scholars prefer more straightforward and transparent models if the captured variation is sufficient. For example, Henry and Leone (2016) criticise a Naive Bayes machine learning algorithm as it performs similarly to word lists while it has more drawbacks as it is more opaque and harder to implement. Second, unlike dictionaries, machine learning algorithms typically require a substantial amount of training data. Third, machine learning models are less accessible than lexicons as they require sufficient programming understanding and computing power.

2.3 Managing Disclosure Sentiment

Managers have an incentive to be ambiguous in disclosing negative information as their compensation is tied to the firm's performance (Hollander et al., 2010; Kothari, Shu, & Wysocki, 2009). Despite the regulation surrounding financial disclosures, it is recognised that disclosures are generally positively biased. For instance, managers withhold bad news (Kothari et al., 2009), provide incomplete information (Hollander et al., 2010), and use more favourable tones than other call participants in conference calls (Loughran & McDonald, 2011). Huang, Zang, and Zheng (2014) document that managers use an abnormally positive tone when they face incentives to artificially boost the perception of investors or when the firm is close to meeting or beating consensus analysts' forecasts. Further, the authors find that disclosure tone manipulation is fruitful as markets react positively to abnormal management optimism. Similarly, Arslan-Ayaydin, Boudt, and Thewissen (2016) show that managers are more inclined to use optimistic language in press releases when their equity-based incentives are more sensitive to changes in the stock price.

Evidence of this so-called sentiment managing is in line with the hypothesis that managers steer clear from words that are perceived as negative by algorithms. Cao et al. (2020) find that managers of firms with higher expected machine disclosure readership express more excitement and positivity in their tone. The authors note that this is consistent with the anecdotal evidence that managers are trained by professionals to influence NLP algorithms with their vocal performances. In addition, they find that managers have started to avoid words that are tagged as negative in the LM Dictionary since its publication in 2011. Consequently, researchers should consider the possibility that firms keep away from negative words in widespread sentiment word lists.

2.4 BERT

Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), or BERT, is a language model that involves a layered structure of bidirectional Transformers. The Transformer (Vaswani et al., 2017) is an encoder-decoder deep neural network architecture. Similarly to recurrent neural networks (RNNs), Transformers are designed to handle sequential information. Transformers have two main advantages over RNNs, such as LSTMs. First, due to the sequential nature of RNNs, information is lost as the sequence gets longer. Transformers have no locality bias as the self-attention architecture processes the entire input sequence at once (Munikar et al., 2019). Second, Transformers are more efficient during training as they are better suited for parallelisation.

BERT uses the Transformer's encoder component to processes the input sequence to generate word embeddings that represent contextual information. The key innovation of BERT is the deeply bidirectional pre-training. Previous models are limited to unidirectional (left-to-right or right-to-left) training. BERT's trained sense of language does not follow a single direction. The model is pre-trained to understand language on two unsupervised tasks simultaneously:

(1) *Hidden word prediction:* To learn to predict hidden words, BERT takes sentences with 15% of the words masked by a `[MASK]`-tags and trains to complete sentences based on the other words.

(2) *Next sentence prediction:* The model is fed two sentences and learns whether the second sentence follows the first sentence or is a random sentence.

BERT is trained on English Wikipedia articles (2,500 million words) and BookCorpus (800 million words)[1]. There are two versions of BERT: BERT-base, with 12 encoder layers and 110 million parameters, and BERT-large, with double the encoder layers and 340 million parameters (Devlin et al., 2019). BERT takes four days to pre-train on 4 to 16 Google Cloud TPUs. This is a one-time process; NLP researchers can then fine-tune the pre-trained model to their needs. Fine-tuning is computationally inexpensive as it only requires minimal architectural changes, making it a suitable transfer learning model. Therefore, BERT may be part of the answer to the questions raised about the negative environmental impact of training language models (e.g. Strubell et al., 2019).

BERT can be fine-tuned to solve diverse downstream tasks that require language understanding. For example, BERT has shown state-of-the-art results on neural machine translation (Zhu et al., 2019), named entity recognition (Lee et al., 2020), and question answering (van Aken et al., 2019). Various researchers demonstrate the performance of BERT on sentiment classification tasks. For example, BERT achieves excellent performances on aspect-based sentiment analysis (Li et al., 2020). Munikar et al. (2019) use BERT to predict the sentiment of one-sentence movie reviews from Rotten Tomatoes on a 1 (very negative) to 5 (very positive) scale. They show that their simple BERT model outperforms many other sophisticated NLP models, such as the BiLSTM.

2.5 FinBERT

Loughran and McDonald (2011) document that sentiment classification is a domain-specific task. Thus, pre-training BERT on financial texts instead of general texts may increase performance on

[1]See `https://github.com/google-research/bert` for the TensorFlow code and pre-trained weights.

financial sequence classification tasks. Yang et al. (2020) introduce FinBERT, a BERT-based NLP model. Unlike BERT, FinBERT is trained on a financial corpus that consists of corporate reports (2.5 billion tokens), earnings call transcripts (1.3 billion tokens), and analyst reports (1.1 billion tokens). The authors show that FinBERT accomplishes accuracy gains in financial sentiment tasks compared to the generic BERT model. For example, FinBERT accurately predicts the tone of 88.7% of the 10,000 manually tagged sentences in the AnalystTone data set (Huang, Zang, & Zheng, 2014), a 5.5% improvement over BERT. This suggests that pre-training on domain-specific texts increases the model's understanding of domain-specific language. This is also highlighted by the successes of other domain-specific BERT adaptations such as BioBERT (Lee et al., 2020).

2.6 Contributions

This study contributes to the finance and accounting textual analysis literature in the following ways. First, I provide insights into the shortcomings of sentiment lexicons compared to FinBERT. Second, I thoroughly evaluate the performance of FinBERT as tone classifier on a large sample of conference calls Q&As. Third, I address the potential presence of managerial sentiment manipulation in a new setting.

3 Methods

3.1 Text Samples

This study uses various forms of financial texts. First, I use annotated sentences from financial news, firm press releases, and microblog messages. My objective with this data is to assess the out-of-sample accuracy of the Harvard IV-4 Dictionary, the LM Dictionary, and FinBERT. Second, I employ a large sample of earnings calls transcripts from U.S. firms. I only use the Q&A section of earnings calls as previous work shows that this part is more informative than the scripted management presentation (Mayew, Sethuraman, & Venkatachalam, 2020).

Pre-processing

Before applying the sentiment algorithms, I carry out several transcript pre-processing steps. First, I remove all text between tag symbols (e.g. "[coughing]"). Second, I convert the text to lowercase. Third, I replace symbols with spaces. Fourth, I remove all non-alpha characters except spaces. Fifth, I require sentences to have at least 5 words to remove insignificant remarks (e.g. "Yeah thanks.").

3.2 Tone Measures

Dictionaries

As a baseline, I measure financial sentiment using two dictionaries. First, I use the Harvard IV-4 Psychosocial Dictionary[2]. This is one of the earliest semantic dictionaries and it is still relatively popular for sentiment classification. This lexicon has 1,563 words in the positive list and 1,892 words in the negative list. Second, I use the LM Dictionary[3] as my main baseline. The word lists contain 354 positive and 2,355 negative words. There are numerous textual tone word lists, but this dictionary is arguably the gold standard of measuring sentiment in financial texts.

I classify sentences as positive (negative) if it contains more words from the positive (negative) word list than from the negative (positive) word list of the respective dictionaries. Sentences with an equal number of positive and negative words are treated as neutral sentences.

FinBERT

FinBERT (Yang et al., 2020) is trained on a large financial corpus that consists of corporate reports (2.5 billion tokens), earnings call transcripts (1.3 billion tokens), and analyst reports (1.1 billion tokens). FinBERT uses the BERT-base architecture, with 12 encoder layers and 110 million parameters. It takes two days to pre-train FinBERT on four Tesla P100 GPUs.

I apply the following additional pre-processing steps on the texts before I feed them to FinBERT:

(1) *Tokenization:* First, I apply WordPiece tokenization using the Transformers package[4] in Python.

(2) *Padding:* The maximum sequence length that I use is 64 tokens. A low proportion of outliers are truncated (0.8%), whereas the bulk of sentences in my sample are padded to the maximum length with `[PAD]`-tokens.

(3) *Add special tokens:* Second, I add `[CLS]`-tokens to separate sequences and `[SEP]`-tokens to separate sentences, as required by BERT.

As recommended by its developers, I use the pre-trained uncased FinBERT model. I fine-tune FinBERT on 10,000 sentences from the AnalystTone data set (Huang, Teoh, & Zhang, 2014). It contains 3,580 positive, 4,590 neutral, and 1,830 negative phrases, randomly selected from analysts reports. This fine-tuning process takes less than one hour on a single GPU. The output

[2]See http://www.wjh.harvard.edu/~inquirer/homecat.htm.
[3]See https://sraf.nd.edu/textual-analysis/resources/.
[4]The Transformers package is available on: https://github.com/huggingface/transformers.

neural network layer of fine-tuned FinBERT returns a vector of logits that I transform into three probabilities using a softmax function. Each sentence is labelled with the class (*positive*, *negative*, or *neutral*) with the highest predicted probability.

Variable Construction

Following previous studies (e.g. Chen et al., 2018; Henry & Leone, 2016), tone in this study is defined in the following way:

$$Tone_j = \frac{Pos_j - Neg_j}{Pos_j + Neg_j}, \tag{1}$$

where, *Pos* (*Neg*) is the number of positive (negative) sentences as measured in the quarterly earnings call *j*. This form ensures that *Tone* is bounded between -1 and 1. I standardise the *Tone* estimates so that their mean is 0 and their standard deviation is 1.

3.3 Research Design

As my interest is in the applicability of FinBERT in finance and accounting research, I emulate typical financial sentiment research methods with these texts. Therefore, I use a panel data analysis with ordinary least squares (OLS) regressions. My panel consists of S&P 500 firms for the 17-year period from 2004 until 2020. I focus on the S&P 500 as the high analyst following of these firms ensures an informative Q&A section of earnings calls (Mayew et al., 2020). I use stock returns (Section 4.2) and changes in output that individual analysts produce (Section 4.3) to measure the underlying sentiment of earnings calls.

Data Collection

I retrieve transcripts of earnings conference calls from Thompson Reuters Street Events[5]. The transcripts are generally well-structured text files, with headers dividing each file into a presentation and a Q&A section. However, a small proportion (roughly 1.5%) of the files cannot be parsed as they contain structural errors. The transcripts contain metadata including the date, time, and firm identifiers. Additionally, I use a parsing algorithm[6] to extract the participant identity in the form of a numeric identifier, name, role (e.g. operator, CEO, analyst), and the company that the speaker represents along with the spoken text. This method yields 26,237 transcripts of 715 unique firms for my analyses.

[5]I thank Stephan Hollander for providing the raw transcript data.
[6]My parsing algorithm, and all other Python scripts that I use in this paper, are available on https://github.com/luukhopman/measuring-financial-tone.

The stock returns that I use in this study to proxy the sentiment of post-earnings market reaction are gathered from the Center for Research in Security Prices (CRSP). My analyst report variables, which I consider the *true* underlying analyst sentiment, is from Institutional Brokers' Estimate System (I/B/E/S). I match the participating analysts in earnings calls to their stock recommendations, price targets, and EPS forecasts. As the merging relies on the names of analysts in the header of the transcript, which are sometimes misspelt, a proportion of analysts remain unmatched to the I/B/E/S files. Data for control variables that I construct are from various sources, including the financial databases Compustat and CRSP. Variables constructed from Compustat and CRSP data are matched to the relevant earnings call by company identifier (GVKEY or Ticker) and quarter. For additional details, see the variables definitions and sources in Appendix Table A1.

Robustness

Following the prior research in the domain (e.g. Chen et al., 2018; Henry & Leone, 2016; Price et al., 2012; Loughran & McDonald, 2011), I include a common set of firm-specific control variables. *Surprise* reflects the unexpected earnings, i.e. the actual earnings per share (EPS) *minus* the consensus EPS forecast. *Size* is a proxy for firm size, defined as the natural logarithm of market capitalisation. Return on assets (*ROA*) refers to the per cent ratio between the quarterly net income and the book value of total assets. The ratio between long-term debt and total assets (*Leverage*) controls for the disclosure differences associated with financial distress. Similarly, *Loss* is a dummy variable that indicates negative net income in the relevant quarter. The capital expenditures scaled by total assets (*Capex*) accounts for differences between growth opportunities. Additionally, I control for the number of words in the Q&A section of the earnings call transcripts (*NumWords*) as an indicator for the quantity of conveyed information (Price et al., 2012).

I use various model specifications to robustify my results. In certain model specifications, I include time fixed effects (i.e. year dummies) to control for macroeconomic trends and firm fixed effects (i.e. firm identifier dummies) to control for time-invariant heterogeneity between firms. Moreover, all the reported t-statistics are based on robust standard errors, clustered at the firm level. The use of robust standard errors prevents violations of the homoscedastic errors assumption of the Gauss-Markov theorem (White, 1980). Clustering standard errors is a common technique to control for model error correlation within entities (i.e. firms).

4 Experimental Setup

In this section, I conduct several experiments to answer the research questions in this study. Section 4.1 introduces the experiment that I use to measure the performance of various *Tone* measures on annotated data (RQ1). Sections 4.2 and 4.3 depict the two experiments that test the statistical (RQ2) and economic significance (RQ3) of using FinBERT rather than the LM Dictionary in a research setting. Last, Section 4.4 outlines the experiment that I conduct to determine the presence of managerial tone manipulation in my sample (RQ4).

4.1 Labelled Financial Phrases

In this first experiment, I examine the performance of sentiment classification models on two data sets with human-labelled financial phrases. I compare the two dictionaries, the Harvard-IV-4 dictionary and the LM financial dictionary, with FinBERT. Specifically, I quantify performance on this task with the accuracy and F1-score.

The aim of this experiment is twofold. First, it shows the performance of FinBERT on annotated test data relative to dictionaries (RQ1). If FinBERT is a viable alternative to the popular LM Dictionary, it should perform significantly better on this task. Second, I exhibit some repeated mistakes caused by the weaknesses of the LM Dictionary using example sentences.

The first annotated set of phrases that I use for this experiment is from the Financial PhraseBank[7] (Malo, Sinha, Korhonen, Wallenius, & Takala, 2014). This data set contains financial expressions from news headlines and press releases of Finnish companies. The statements are tagged as either *positive*, *negative*, or *neutral* by annotators with sufficient financial knowledge (Malo et al., 2014). I decompose the data sets based on the level of agreement between analysts. Starting with the 2,264 sentences that all 16 annotators label identically, I subsequently expand the selection to statements with at least 75% (3,453 sentences) and 50% (4,846 sentences) agreement between annotators.

Thereafter, I use annotated phrases from the Financial Opinion Mining and Question Answering (FiQa) data set that was part of the WWW'2018 Open Challenge[8] (Maia et al., 2018). This data contains 436 financial news headlines and 675 Twitter messages, labelled with a continuous sentiment score ranging from -1 (very pessimistic) to 1 (very optimistic). I map this value to a categorical variable: *negative* (smaller than -0.4), *neutral* (between -0.4 and 0.4), and *positive*

[7]The Financial PhraseBank can be found here https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10.
[8]See https://sites.google.com/view/fiqa/home.

(larger than 0.4). The thresholds are arbitrary, but my conclusions are insensitive to adjustments. In this experiment, I differentiate between the headlines and posts in the FiQa data set to examine the effect of various text origins.

4.2 Market Reaction

I explore the predictive power of the two *Tone* measures in an event study with stock returns. The intuition behind this experiment is that one would expect a positive market reaction after a positive quarterly conference call (Huang, Teoh, & Zhang, 2014; Davis et al., 2012; Loughran & McDonald, 2011). Multiple studies use the market reaction to disclosure tone and successfully show that the LM Dictionary is superior at capturing financial sentiment compared to the Harvard IV-4 Dictionary (Henry & Leone, 2016; Price et al., 2012; Loughran & McDonald, 2011). I apply the same method to further explore the performance differences between the LM Dictionary and FinBERT.

This experiment has two objectives. First, this experiment provides insights on the effect of measuring financial sentiment with FinBERT rather than with the LM Dictionary in empirical financial research on the statistical power (RQ2). Second, similarly, it gives an estimate of the economic impact (RQ3). Overall, the results will indicate the implications of using FinBERT in future research. Another reason to explore the capacity of FinBERT beyond labelled phrases is that relying entirely on human-labelled data may introduce biases (Jegadeesh & Wu, 2013).

I calculate cumulative abnormal returns (CAR) to isolate the stock price reaction surrounding the earnings conference call. Following the *Market Adjusted Model*[9], I define abnormal returns (AR_i) as the difference between the raw stock return (R_i) for firm i on day t and the CRSP value-weighted market return ($R_{m,t}$) (i.e. the weighted-average of stock returns by firm market capitalisation):

$$AR_i = R_i - R_{m,t}. \tag{2}$$

I cumulate the abnormal returns over a window from one day before the earnings conference call until one day after the call[10]. This is a routinely used short-term CAR window (e.g. Henry & Leone, 2016; Price et al., 2012).

$$CAR[-1,+1] = \sum_{t=-1}^{1} AR_i. \tag{3}$$

[9]For extensive information about event studies in financial research, abnormal returns, and the Market Adjusted Model, I refer the reader to this paper: MacKinlay (1997).

[10]I experiment with different event windows including [0,+1], [-1,+3], and [0,+3], the results are all comparable (untabulated).

I analyse the effect of conference Q&A tone on short-window CARs on the firm level. The linear regression model that I use has the following form:

$$\begin{aligned} CAR[-1,+1]_j = &\gamma_0 + \gamma_1 Tone_j^m + \gamma_2 NumWords_j + \gamma_3 Surprise_j + \gamma_4 Size_j \\ &+ \gamma_5 ROA_j + \gamma_6 Leverage_j + \gamma_7 Loss_j + \gamma_8 Capex_j \\ &+ d_t + d_i + \varepsilon_j. \end{aligned} \tag{4}$$

Here, j indexes earnings conference calls. *Tone* refers to the tone in the Q&A component of the earnings calls with $m \in \{LM, FinBERT\}$. *NumWords* controls for the informativeness of the Q&A, proxied by the number of words in the Q&A transcript, d_t and d_i reflect year and firm dummies respectively. See Section 3.3 for the descriptions of the other controls. All variables definitions and data sources are available in Table A1.

Financial Tone Trading Strategy

Besides the economic impact of using FinBERT in finance and accounting research, I am also interested in its effect on algorithmic trading strategies based on disclosure sentiment. Accordingly, I divide my sample into cross-sections of high and low sentiment portfolios for both $Tone^{LM}$ and $Tone^{FinBERT}$. The high (low) sentiment portfolio contains earnings calls observations in the highest (lowest) quartile of the respective tone measures. For example, the high $Tone^{LM}$ portfolio consists of the 25% sample observations with the highest $Tone^{LM}$ score. This setting enables me to test the differences between low and high portfolios of both tone measures. One would expect that the difference between the high and low portfolio returns to be higher as the accuracy of the *Tone* measure improves, as this indicates that the measure better reflects the market interpretation of the underlying sentiment. Specifically, I analyse the CARs differences between the high and low tone portfolio in the short term ([-1,+1]) and during the earnings drift ([+2,+60], see Price et al. (2012)) with t-tests and Wilcoxon rank-sum tests. I then show the returns of a sentiment trading strategy based on *Tone* determined by the LM Dictionary or FinBERT over the sample period. The strategy entails the purchase of a stock after a positive earnings call and the short selling of a stock after a negative earnings call. Thus, the most precise *Tone* proxy should produce the highest returns.

4.3 Analyst Sentiment

Next, I investigate whether FinBERT can better capture the tone of analysts in earnings conference calls compared to the LM Dictionary. More specifically, I use the post-call output of analysts to proxy their *true* sentiment of the earnings calls. A higher correlation between measured sentiment in earnings calls and the sentiment of the analysts' output suggests that the measure is a better approximation of how analysts interpret the earnings call. Unlike the stock return model, this experiment is on the individual level, that of the analyst, and therefore more challenging.

I label analyst tone by utilising their revisions in individual analysts output reports that reflect their contemporary view on the firm. Following Chen et al. (2018), I use a 20-day revision window after the day of the earnings call to provide sufficient time for analysts to produce said output. The three analyst outputs that I use are as follows:

- *Analyst Recommendations:* I compare the analysts stock recommendation level (*sell*, *underperform*, *hold*, *buy*, or *strong-buy*) that the analyst reports after the earnings call to his or her previous recommendation level. I map a one-level stock rating upgrade as +1, a three-level downgrade as -3 and so on.
- *Analyst Price Targets:* I measure changes in price targets (i.e. the analyst's projection of the future stock price) by calculating the difference between the new price target in the revision window to the previous price target set by the analyst. I scale the change in price target by the previous price target.
- *Analyst EPS Forecast:* Similar to the price target measure, I consider the change in the EPS Forecast after the earnings conference call and scale it by the previous EPS forecast.

I use the three variations of analyst output as the dependent variable in a regression model with the same control variables as in Equation 4:

$$\begin{aligned} AnalystOutput_j = &\alpha_0 + \gamma_1 AnalystTone_j^m + \gamma_2 NumWords_j + \gamma_3 Surprise_{i,t} \\ &+ \gamma_4 Size_j + \gamma_5 ROA_j + \gamma_6 Leverage_j + \gamma_7 Loss_j \\ &+ \gamma_8 Capex_j + d_t + d_i + \varepsilon_j, \end{aligned} \tag{5}$$

where *AnalystOuput* refers to the three output metrics (*Recommendation*, *PriceTarget*, and *EPSForecast*). *AnalystTone* reflects the individual analyst's *Tone* during the Q&A of conference call j with $m \in \{LM, FinBERT\}$.

4.4 Managerial Tone Manipulation

Prior literature suggests that managers adapt their textual and vocal performances to NLP models. Cao et al. (2020) are among the first to research this. They find a significant decrease in the use of negative sentiment measured by the LM financial dictionary after its publication in 2011. The authors show that this trend is not present for sentiment measured by the Harvard IV-4 dictionary. To better understand this phenomenon, I carry out an experiment to test whether this potential trend of avoiding negative words is measurable in my sample.

To answer RQ4, I construct an experiment to test whether there is a similar trend observable for the managers in my data set. Conference calls have the unique characteristics of featuring both firm insiders and outsiders. Firm insiders have an incentive to circumvent negative language to mislead NLP models while outsiders have no apparent reason to choose their words this carefully.

Specifically, I design a difference-in-difference (DID) regression model. In this model, I place the senior managers in earnings conference calls in the treated group, and analysts in the same calls in the control group. Unlike managers, analysts have no obvious incentive to avoid bad press and to circumvent NLP algorithms. I select the Q&A participants and divide them into two groups. First, the management, specifically the C-level executives (e.g. CEO, CFO), tagged by a regular expression. Second, the analysts, whom I identify by their speaker role in the transcripts. In my sample, the analysts speak 1.8 million sentences, totalling 35.1 million words. Whereas managements speak 4.2 million sentences and 85.3 million words in the earnings calls Q&As. I use the publication date (6 January 2011) of Loughran and McDonald (2011) as the cutoff between the pre- and post-intervention period. This is in line with Cao et al. (2020), who use the publication date as the instrumental effect in their study.

$$\begin{aligned} NegativeWords_j =& \gamma_0 + \gamma_1 Treated_i \times Post_t + \gamma_2 Treated_i + \gamma_3 Post_t + \gamma_3 NumWords_j \\ &+ \gamma_4 Surprise_j + \gamma_5 Size_j + \gamma_6 ROA_j + \gamma_7 Leverage_j \\ &+ \gamma_8 Loss_j + \gamma_9 Capex_j + d_t + d_i + \varepsilon_j, \end{aligned} \tag{6}$$

where *NegativeWords* is a count of the negative words spoken by the conference call participants. *Treated* equals one for managers and zero for analysts. *Post* is a dummy indicating whether the date of the conference call is before or after the publication of the LM Dictionary on 6 January 2011. *NumWords* is included to control for the number of words spoken by the participant group.

The regular control variables, year (d_t) and firm (d_i) dummies are also included.

Under the hypothesis that managers avoid negative words in the LM Dictionary, one would expect the coefficient γ_1 to be negative. I also expect a negative sign for γ_2 as that would be consistent with previous findings that suggest that managers are more optimistic in conference calls than analysts (for example, see Figure 2).

5 Results

5.1 Descriptive Statistics

Table 1 reports the summary statistics for the main variables in this study. The unstandardized mean (median) values of $Tone^{LM}$ and $Tone^{FinBERT}$ are 0.154 (0.172) and 0.302 (0.322) respectively. Thus, earnings calls tend to be more positive than negative. The mean of $CAR[-1,+1]$ of 0.098% indicates a slightly positive short-term market reaction surrounding the average earnings call. On average, the reported EPS are higher than the consensus EPS in my sample, as suggested by the positive mean of *Surprise*. 8.2% of the firm-quarter observations in my sample report a quarterly loss.

Table 1: Summary statistics of variables. All values are before standardising. The sample includes firms in the S&P 500 from 2004 to 2020. Variable definitions are reported in Table A1.

	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)	N
$Tone^{LM}$	0.154	0.242	0.000	0.172	0.326	21,966
$Tone^{FinBERT}$	0.302	0.272	0.126	0.322	0.500	21,966
$CAR[-1,+1]$	0.098	6.074	-2.893	0.110	3.231	21,966
Surprise	1.381	16.842	0.885	1.331	1.821	21,966
Size	9.792	1.118	9.037	9.669	10.442	21,966
ROA	1.609	2.453	0.610	1.417	2.539	21,966
Leverage	0.233	0.165	0.115	0.213	0.323	21,966
Loss	0.082	0.274	0.000	0.000	0.000	21,966
Capex	2.367	2.777	0.592	1.479	3.137	21,966

Pearson pairwise correlations of selected variables indicate that both *Tone* measures are positively correlation with $CAR[-1,+1]$, *Surprise*, *Size*, *ROA*, and *Leverage* (see Table 2). *Loss* and *Capex* are negatively related to *Tone* in earnings calls. Given that the correlations between independent variables are relatively low, multicollinearity is unlikely to be problematic in my models.

Table 2: Pearson correlation coefficients between variables. The sample includes firms in the S&P 500 from 2004 to 2020. Variable definitions are reported in Table A1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) $Tone^{LM}$	1.000								
(2) $Tone^{FinBERT}$	0.710	1.000							
(3) $CAR[-1,+1]$	0.114	0.163	1.000						
(4) *Surprise*	0.036	0.018	0.005	1.000					
(5) *Size*	0.142	0.173	-0.009	0.020	1.000				
(6) *ROA*	0.106	0.057	0.046	0.073	0.174	1.000			
(7) *Leverage*	0.027	0.024	-0.015	0.013	-0.042	-0.094	1.000		
(8) *Loss*	-0.090	-0.026	-0.054	-0.081	-0.164	-0.486	0.081	1.000	
(9) *Capex*	-0.032	-0.047	-0.003	0.047	-0.013	0.029	0.046	0.033	1.000

Figure 1 displays the Q&A sentiment in earnings conference calls over the sample period as estimated by the LM Dictionary and FinBERT. The correlation between the two measures is 0.710. Unsurprisingly, both *Tone* measures are at their lowest during the financial crisis of 2008. Towards the end of the sample period another dip in *Tone* is noticeable, with worries about the COVID-19 virus arising in earnings calls (see Hassan, Hollander, Van Lent, Schwedeler, and Tahoun (2020)).

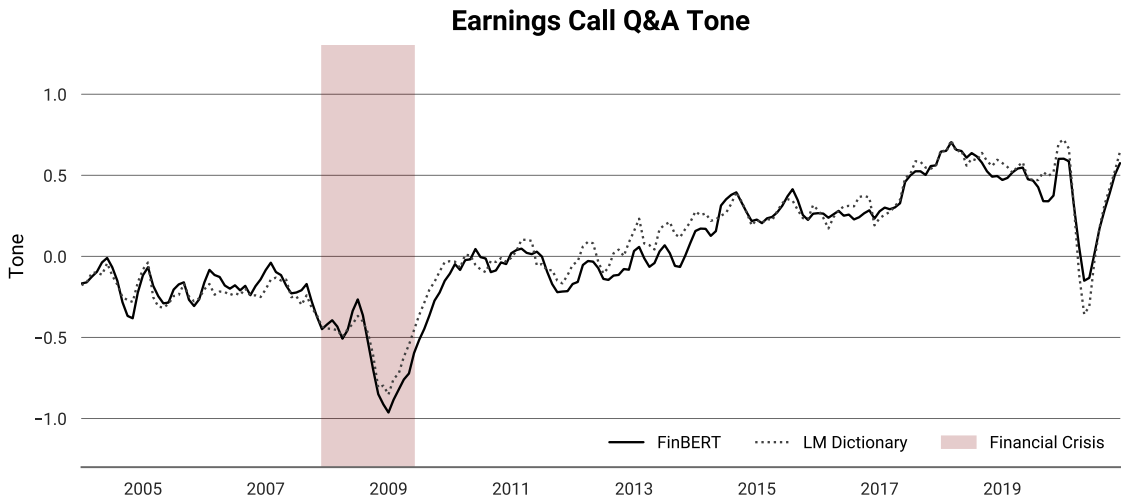

Figure 1: Tone in earnings conference calls Q&As. The figure shows the three-month moving average. *Tone* measures are standardised to have a mean of 0 and a standard deviation of 1.

Figure 2 shows that corporate insiders (management) are more optimistic than the outsiders (analysts) in earnings calls Q&As. The correlation between tone estimated by the LM Dictionary and FinBERT is higher for management *Tone* (0.663) than for analyst *Tone* (0.492).

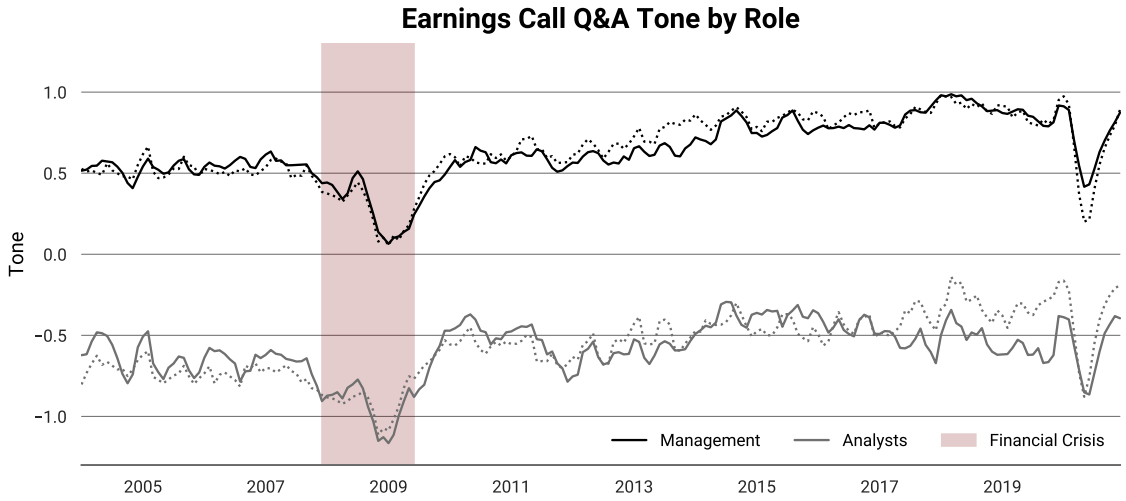

Figure 2: Management and analysts *Tone* in earnings conference calls Q&As. The solid lines represent *Tone* measured by FinBERT, the dotted lines represent *Tone* measured by the LM Dictionary. The figure shows the three-month moving average.

5.2 Labelled Financial Phrases

The results of the first experiment (see Section 4.1) are reported in Table 3. The Harvard IV-4 dictionary achieves an accuracy of 54.59% and an F1-score of 54.87% on the statements with 100% agreement in the Financial PhraseBank test set. The LM Dictionary performs significantly better on this task, with an accuracy of 65.02% and an F1-score of 59.96%. This finding is in line with earlier findings that financial-tailored sentiment dictionaries outperform general sentiment dictionaries. FinBERT's accuracy (F1-score) of 91.65% (91.36%) is significantly better than that of the dictionaries.

Table 3: Performance on annotated financial sentences.

	IV-4 Harvard Dictionary		LM Dictionary		FinBERT	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Financial PhraseBank						
100% Agreement	54.59%	54.87%	65.02%	59.96%	91.65%	91.36%
75% Agreement	54.30%	54.98%	65.77%	61.85%	87.69%	87.26%
50% Agreement	52.79%	53.18%	62.57%	59.29%	79.24%	78.42%
FiQa						
Headlines	52.98%	54.49%	59.17%	58.04%	67.20%	67.74%
Posts	39.70%	38.97%	41.33%	39.39%	50.22%	50.08%

The confusion matrices in Figure 3 show the superior performance of FinBERT on the three classes. It should be noted is that FinBERT has some struggles with detecting positive tone in phrases. The Harvard IV-4 Dictionary's poor performance on negative phrases stands out, as more than 80% of these are misclassified as neutral or positive sentences. The LM Dictionary

tends to misclassify sentences with sentiment as neutral sentences. Interestingly, opposing the general-purpose dictionary, the worst performance is on the phrases with a positive label.

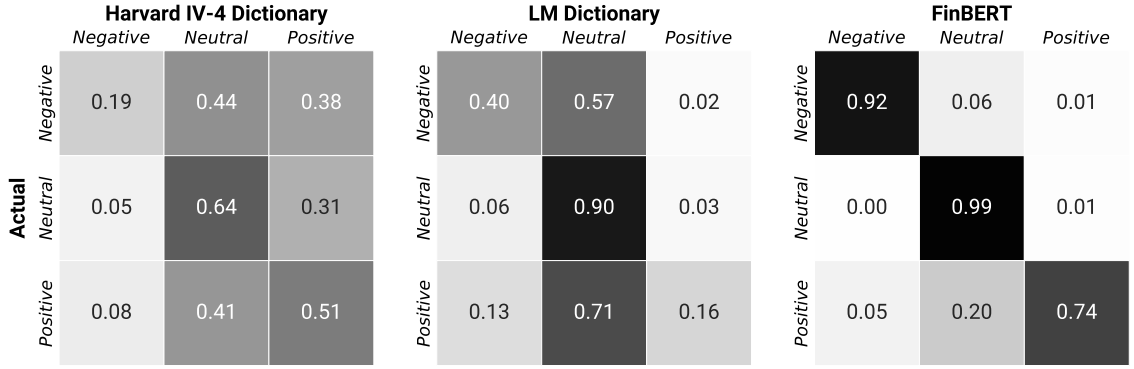

Figure 3: Financial PhraseBank confusion matrix. Normalised by actual labels. This figure only concerns the statements with 100% agreement. The confusion matrix for sentences with at least 50% agreement is displayed in Appendix Table A1.

FinBERT is also the top performer on the FiQa labelled sentences (Figure 4). This task seems harder as the accuracy on the headlines (67.20%) and Twitter posts (50.22%) is much lower compared to its accuracy on the Financial Phrasebank. Again, the finance-specific dictionary scores better than the Harvard IV-4 Dictionary in both categories. However, the accuracy on the Tweets is only slightly above chance-level for both models (Appendix Table A2). The relatively poor performance on Twitter messages is understandable as none of the models are fitted on social media language.

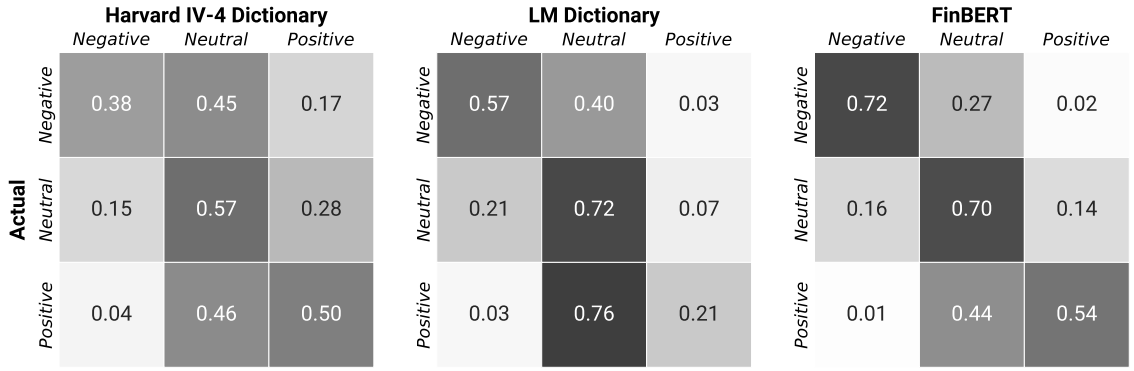

Figure 4: FiQa news headlines confusion matrix. Normalised by actual labels. The confusion matrix for FiQa Twitter posts is displayed in Appendix Table A2.

Overall, the results of this experiment suggest that FinBERT is the superior financial sentiment classifier in this comparison. I present a selection of example sentences from the two labelled data sets to highlight some of the frequent errors that are caused by using the LM Dictionary in financial sentiment classification tasks:

Example 1 "The announced restructuring will significantly decrease the company's indebtedness."

Label: *Positive* LM Dictionary: *Negative* FinBERT: *Positive*

Example 2 "In the second quarter of 2010, Raute's net loss narrowed to EUR 123,000 from EUR 1.5 million in the same period of 2009."

Label: *Positive* LM Dictionary: *Negative* FinBERT: *Positive*

Example 3 "US dollar wipes out sales gains for SABMiller."

Label: *Negative* LM Dictionary: *Positive* FinBERT: *Negative*

Example 4 "I am not optimistic about $AMZN, both fundamentals and charts look like poopoo this quarter."

Label: *Negative* LM Dictionary: *Positive* FinBERT: *Negative*

These examples underline the importance of context in linguistic classification. Example 1 portrays the most common type of error. The token *restructuring* is negative in the LM Dictionary, but it is crucial to factor in whether the context is optimistic or pessimistic. The second example also shows this, without considering the context that it *narrowed*, the token *loss* is negative. Further, word lists cannot assess numeric sentiment (EUR 123,000 is lower than EUR 1.5 million). It should be noted that the current implementation of FinBERT also disregards numbers, but it is still able to assign the correct label in this case. Example 4 reveals two other shortcomings of dictionaries. (1) Informal speech that is prevalent on social media are not in the word lists. This also applies to financial jargon, such as *bullish* (i.e. optimistic) and *bearish* (i.e. pessimistic). (2) A simple negation "*not* optimistic" cannot be interpreted by a unigram word list. Altogether, the examples show the weaknesses of dictionaries and FinBERT shows that it can handle complex language.

In Appendix Tables A3 and A4, I report the positive and negative words in the LM Dictionary with the highest and lowest accuracy in this experiment. It shows the large accuracy differences between words. For example, the sentences with word *improved* are classified with 100% accuracy. Whereas sentences containing the negative word *bridge* were actually 60% positive and 40% neutral.

5.3 Market Reaction

Table 4 reports the results of estimating the market reaction OLS regression models. For both *Tone* measures, I report two regression specialisations to robustify my results. *Tone* is a statisti-

cally significant predictor of $CAR[-1,+1]$ in all four model specifications. Thus, more optimism in earnings calls is correlated with higher contemporary stock returns.

One standard deviation in $Tone^{LM}$ results in a 1.088 per cent point increase in $CAR[-1,+1]$ (Column (2)). The economic magnitude of $Tone^{FinBERT}$ on the dependent variable is larger: one standard deviation change results in a 1.528 per cent point change. This indicates that the LM Dictionary underestimates the economic magnitude with 44.0% with respect to FinBERT. Both tone measures, however, lead to the same inference (i.e. optimistic earnings call tone is related to a higher CAR) at the 1% statistical confidence level.

It should be noted that the adjusted R^2 is low in all regressions. Although the explained variance in stock return is higher in the regressions with $Tone^{FinBERT}$, the adjusted R^2 is only 6.0%. Or as Loughran and McDonald (2011) note, our abilities to justify stock returns are still limited.

Table 4: Market reaction. This table reports the results of estimating Equation 4. Variable definitions are reported in Table A1. The t-statistics based on robust standard errors clustered by firm are reported in parentheses. ***, **, and * indicate a two-tailed test significance level at 1%, 5%, and 10%, respectively.

	$CAR[-1,+1]$			
	(1)	(2)	(3)	(4)
$Tone^{LM}$	0.715***		1.088***	
	(13.74)		(15.72)	
$Tone^{FinBERT}$		1.026***		1.528***
		(17.75)		(19.42)
NumWords	-0.000***	-0.000***	-0.000***	-0.000***
	(-5.31)	(-3.93)	(-4.99)	(-4.40)
Surprise	-0.001	-0.001	0.003	0.004
	(-0.32)	(-0.28)	(0.84)	(0.95)
Size	-0.181***	-0.259***	-1.632***	-1.703***
	(-4.47)	(-5.71)	(-10.12)	(-10.46)
ROA	0.063**	0.069**	0.092**	0.095**
	(2.19)	(2.36)	(2.40)	(2.46)
Leverage	-0.482*	-0.502*	0.566	0.524
	(-1.91)	(-1.89)	(0.92)	(0.85)
Loss	-0.882***	-1.041***	-0.972***	-1.080***
	(-4.04)	(-4.76)	(-3.92)	(-4.38)
Capex	0.000	0.011	0.041**	0.039*
	(0.04)	(0.86)	(2.05)	(1.95)
Year FE	No	No	Yes	Yes
Firm FE	No	No	Yes	Yes
Adjusted R^2	0.018	0.032	0.039	0.060
Observations	21,967	21,967	21,948	21,948

Financial Performance of Tone Trading Strategy

Figure 5 illustrates the CAR differences between the high and low *Tone* portfolios. The high portfolio of $Tone^{FinBERT}$ and $Tone^{LM}$ both outperform the baseline portfolio. Both show a sharp increase in abnormal return on the day of the quarterly earnings call (day 0) and the day after the call (day 1). The post-earnings drift after day 1 is weak compared to the contemporary reaction. The mean CAR of the high $Tone^{FinBERT}$ portfolio is 1.89%, whereas the mean conference call in the high $Tone^{LM}$ portfolio is met with a CAR of 1.36% after 60 days. The low portfolios of $Tone^{LM}$ and $Tone^{FinBERT}$ both have negative average CARs, -0.70% and -0.16% respectively. Again, this indicates that FinBERT is more powerful at detecting positive and negative sentiment in financial texts.

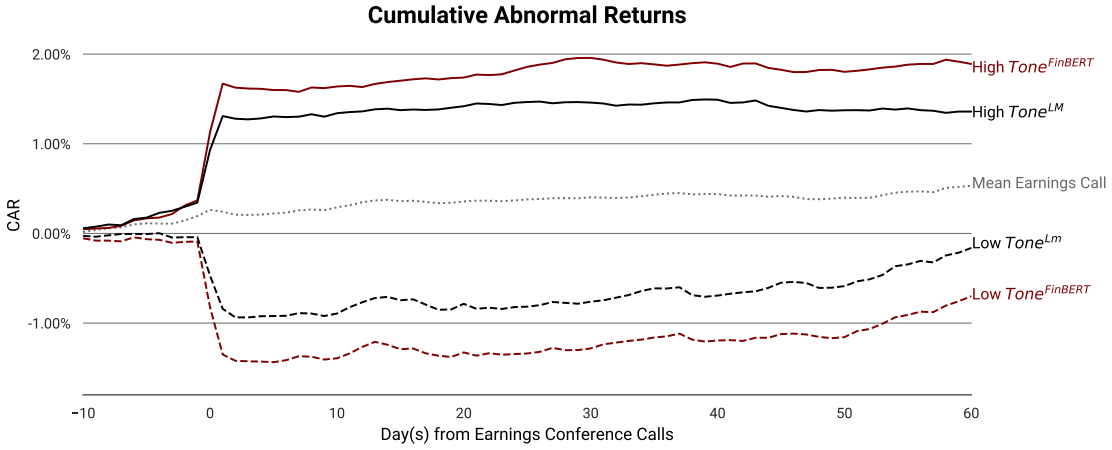

Figure 5: Cumulative abnormal returns (CARs), by low (Q1) and high (Q4) *Tone* quartiles. Abnormal returns, calculated by the adjusted market model, are cumulated from 10 days prior to the earnings calls until 60 days after the earnings call.

I analyse the CAR means and median differences between the high and low *Tone* quartiles using t-tests and the z-statistics of the Wilcoxon rank-sum tests respectively. Table 5 shows the high-minus-low means and medians of the three-day CARs surrounding the earnings call are significantly larger for the Q4 quartile than the Q1 quartile for both *Tone* measures. The mean earnings drift CARs ([+2,+60]) is higher for the Q1 *Tone* portfolios than for the Q4 portfolios, although these differences are statistically insignificant for the medians. However, the positive $CAR[+2, +60]$ for the first quartile may indicate that markets overreact to negative tones in earnings calls and partially revert after the initial reaction.

Table 5: Test of differences between means and medians, by $Tone^{LM}$ and $Tone^{FinBERT}$ quartiles. Variable definitions are reported in Table A1. Test statistics are in parentheses. ***, **, and * indicate a two-tailed test significance level at 1%, 5%, and 10%, respectively.

		$CAR[-1,+1]$		$CAR[+2,+60]$	
		$Tone^{LM}$	$Tone^{FinBERT}$	$Tone^{LM}$	$Tone^{FinBERT}$
Q1 (Low)	Mean	-0.810	-1.275	0.683	0.655
	Median	-0.584	-0.905	0.454	0.445
Q2	Mean	-0.101	-0.219	0.325	0.232
	Median	-0.034	-0.195	0.492	0.408
Q3	Mean	0.305	0.529	0.092	0.060
	Median	0.296	0.456	0.476	0.408
Q4 (High)	Mean	1.022	1.375	0.052	0.222
	Median	0.875	1.182	0.366	0.552
Mean Q4-Q1		1.831^{***}	2.650^{***}	-0.632^{***}	-0.433^{*}
t-statistic		(16.00)	(22.79)	(-2.70)	(-1.86)
Wilcoxon rank-sum test					
Median Q4-Q1		1.459^{***}	2.087^{***}	-0.088	0.107
z-statistic		(17.30)	(23.85)	(-1.06)	(-0.32)

Consequently, the most profitable trading strategy is to long the high $Tone^{FinBERT}$ portfolio and short-selling the low $Tone^{FinBERT}$ portfolio. To further illustrate this, I analyse the returns for the trading strategy that shorts the low quantile of the *Tone* measure and longs the high quantile. I cannot differentiate between the ex-ante and ex-post earnings calls returns on the call date. Therefore, I use the conservative approach of only including the returns of the trading day after the conference call, even though the mean day 0 return is higher than the day 1 return. Figure 6 shows the returns of the FinBERT and LM Dictionary long-short portfolios respectively. Both high-minus-low *Tone* portfolios cumulate higher returns than the benchmark. However, the FinBERT long-short portfolio has significantly higher returns than the LM long-short portfolio. This underlines the positive economic impact of measuring financial text sentiment with FinBERT when it comes to algorithmic trading.

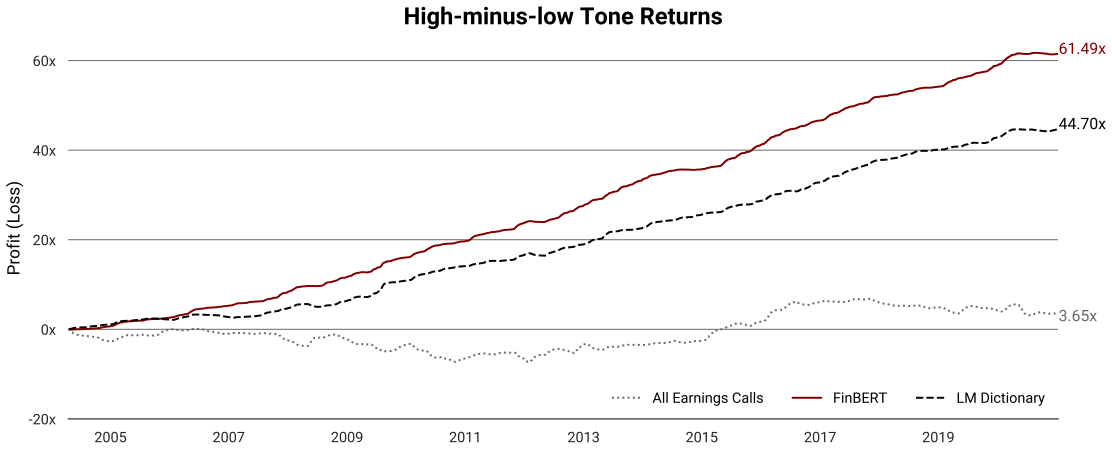

Figure 6: High-minus-low portfolios returns. The cumulative return on the trading day after the earnings call over the sample period. Long positions in stocks with high *Tone* earnings calls, short positions in stocks with low *Tone* earnings calls. The baseline is the aggregate return for all earnings calls. Note: transaction costs and other market frictions are ignored.

5.4 Analyst Sentiment

The results of the analyst output experiment are reported in Table 6. The first two columns show that both $Tone^{LM}$ and $Tone^{FinBERT}$ are positively correlated to recommendations at the 10%- and 1%-level respectively. The difference in statistical certainty is the first evidence for a statistical impact of using FinBERT that I observe in this study. The coefficients of the two tone measures in the *PriceTarget* columns are both statistically significant at the 1-% level. Interestingly, the coefficient of $Tone^{LM}$ in Column (5) is not statistically different from zero, whereas the coefficient of $Tone^{FinBERT}$ is statistically positive with at least 99% certainty. This indicates that the LM Dictionary's measurement error may lead to failing to reject the null hypothesis. In other words, the measurement error can lead to false negatives (type II errors). To conclude, the LM Dictionary underestimates the economic magnitude of the analyst's *Tone* on various analyst outputs.

Table 6: Analyst output. This table reports the results of estimating Equation 5. Variable definitions are reported in Table A1. The *t*-statistics based on robust standard errors clustered by firm are reported in parentheses. ***, **, and * indicate a two-tailed test significance level at 1%, 5%, and 10%, respectively.

	Analyst Output					
	Recommendation		*PriceTarget*		*EPSForecast*	
	(1)	(2)	(3)	(4)	(5)	(6)
$AnalystTone^{LM}$	0.032*		0.530***		0.003	
	(1.74)		(6.35)		(0.85)	
$AnalystTone^{FinBERT}$		0.099***		0.859***		0.008***
		(5.73)		(10.78)		(4.02)
NumWords	-0.001**	-0.001**	-0.005***	-0.004**	-0.000*	-0.000
	(-2.54)	(-2.25)	(-2.65)	(-2.49)	(-1.66)	(-1.59)
Surprise	0.005***	0.006***	0.348	0.349	-0.012**	-0.012**
	(3.17)	(3.43)	(1.25)	(1.26)	(-2.11)	(-2.11)
Size	-0.133**	-0.140**	2.634***	2.606***	0.099***	0.099***
	(-2.14)	(-2.27)	(2.68)	(2.66)	(4.22)	(4.20)
ROA	-0.008	-0.008	0.137	0.137	-0.027**	-0.027**
	(-1.06)	(-1.01)	(1.13)	(1.13)	(-2.14)	(-2.14)
Leverage	-0.101	-0.110	6.515**	6.462**	0.206	0.206
	(-0.31)	(-0.34)	(2.13)	(2.12)	(1.28)	(1.28)
Loss	-0.134	-0.129	0.561	0.579	0.050	0.050
	(-1.52)	(-1.49)	(0.92)	(0.96)	(0.71)	(0.71)
Capex	-0.007	-0.005	-0.369***	-0.365***	-0.014	-0.014
	(-0.64)	(-0.48)	(-3.39)	(-3.33)	(-1.09)	(-1.08)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.006	0.011	0.066	0.067	0.088	0.088
Observations	5,051	5,051	55,831	55,831	100,694	100,694

Effect of Sample Size

Despite the evidence that the LM Dictionary underestimates the economic significance, it captures enough variation to result in statistical significance in the majority of the settings that I explore. The large sample that I have at my disposal likely plays a role in this. Not all financial NLP research has access to a large sample. Therefore, I experiment with subsets of my sample. Specifically, I randomly select 2,500 observations and then re-estimate the regression equation 500 times. The results in Table 7 show that $Tone^{LM}$ has a higher probability of resulting in type II errors on smaller samples. Researchers should therefore consider the possibility of false negatives if they use the LM Dictionary on a small sample.

Table 7: Summary statistics of analysts output regression with bootstrapping. I resample 2,500 observations from the data 500 times and estimate Equation 5. $p < 0.10$, $p < 0.05$ and, $p < 0.01$ refer to the percentage of statistically significant positive effects at the respective levels (two-sided). Variable definitions are reported in Table A1.

		Mean	Positive Sign	$p < 0.10$	$p < 0.05$	$p < 0.01$
Recommendation	$Tone^{LM}$	0.032	94.0 %	21.0%	9.8%	0.8%
	$Tone^{FinBERT}$	0.096	100.0%	99.4%	96.4%	80.4%
PriceTarget	$Tone^{LM}$	0.503	93.4%	46.4%	34.6%	13.6%
	$Tone^{FinBERT}$	0.840	98.4%	82.6%	73.2%	50.8%
EPSForecast	$Tone^{LM}$	0.002	58.6%	4.8%	2.4%	0.4%
	$Tone^{FinBERT}$	0.008	74.4%	14.4%	7.4%	1.0%

5.5 Managerial Tone Manipulation

Table 8 reports the result of estimating Equation 6. The variable of interest is the interaction term between *Treated* and *Post*. In all the model specifications, the sign of the interaction term is significantly negative at the 1% level, indicating that managers use relatively fewer negative words in the LM Dictionary since its publication in 2011. As expected, the negative coefficient of *Treated* suggests that managers use less negative language in conference calls than analysts. The negative sign of the *Post* coefficients in Columns (1) and (2) is an indication that conference calls are more optimistic since 2011. This is not surprising given that the sample period before 2011 includes a span of severe financial recession. In Columns (3) and (4), the year fixed effects subsume the *Post* variable due to collinearity. The significant coefficients of the control variables are also consistent with expectations. *Ceteris paribus*, earnings calls that are longer or hosted by a firm that reported losses contain more negative words. Those of firms with higher earnings or a higher earnings surprise are less negative.

Table 8: Difference-in-difference OLS regression with *NegativeWords* as the dependent variable. Variable definitions are reported in Table A1. The t-statistics based on robust standard errors clustered by firm are reported in parentheses. ***, **, and * indicate a two-tailed test significance level at 1%, 5%, and 10%, respectively.

	NegativeWords			
	(1)	(2)	(3)	(4)
Treated × *Post*	-0.647**	-0.629**	-0.584**	-0.529*
	(-2.20)	(-2.16)	(-2.01)	(-1.87)
Treated	-7.600***	-7.570***	-7.416***	-7.216***
	(-24.55)	(-24.65)	(-24.35)	(-25.60)
Post	-2.851***	-2.753***	4.015***	3.931***
	(-15.35)	(-13.78)	(11.13)	(16.98)
NumWords	0.009***	0.009***	0.009***	0.009***
	(54.26)	(55.55)	(55.44)	(66.36)
Surprise		-0.006	-0.004	-0.003
		(-0.93)	(-0.67)	(-0.45)
Size		-0.173	-0.005	-1.145***
		(-1.04)	(-0.03)	(-5.96)
ROA		-0.173***	-0.155***	-0.083***
		(-3.54)	(-3.24)	(-2.75)
Leverage		-0.100	0.490	1.264
		(-0.11)	(0.52)	(1.13)
Loss		1.447***	1.393***	1.087***
		(3.56)	(3.49)	(3.60)
Capex		-0.180***	-0.204***	0.017
		(-4.27)	(-4.74)	(0.74)
Year FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
Adjusted R^2	0.586	0.590	0.600	0.670
Observations	44,582	44,582	44,582	44,581

I also estimate Equation 6 with *PositiveWords* as the dependent variable (see Table 9). The coefficient of the $Treated_i \times Post_t$ term is statistical significantly positive in all four model specifications. Overall, the results of this experiment support the hypothesis that managers adjust their vocabulary in corporate disclosures to the LM word lists.

Table 9: Difference-in-difference OLS regression with *PositiveWords* as the dependent variable. Variable definitions are reported in Table A1. The t-statistics based on robust standard errors clustered by firm are reported in parentheses. ***, **, and * indicate a two-tailed test significance level at 1%, 5%, and 10%, respectively.

	PositiveWords			
	(1)	(2)	(3)	(4)
Treated × *Post*	4.111***	4.143***	4.101***	4.165***
	(8.50)	(8.56)	(8.48)	(8.77)
Treated	6.631***	6.745***	6.650***	6.788***
	(13.33)	(13.80)	(13.56)	(15.31)
Post	1.878***	1.332***	-16.389***	-0.835**
	(11.01)	(6.18)	(-33.51)	(-2.34)
NumWords	0.014***	0.014***	0.014***	0.014***
	(55.40)	(56.23)	(56.63)	(66.87)
Surprise		0.012***	0.012***	0.002
		(5.01)	(5.00)	(0.93)
Size		0.793***	0.587**	0.869***
		(3.35)	(2.35)	(3.00)
ROA		0.254***	0.251***	0.090**
		(3.17)	(3.16)	(2.48)
Leverage		0.351	-0.714	1.777
		(0.28)	(-0.55)	(1.07)
Loss		-0.205	-0.377	-0.635**
		(-0.39)	(-0.72)	(-2.12)
Capex		-0.162**	-0.136*	-0.131***
		(-2.22)	(-1.87)	(-3.56)
Year FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
Adjusted R^2	0.788	0.790	0.792	0.835
Observations	44,582	44,582	44,582	44,581

I leave the question whether there is a direct causal relationship between the publication of the LM Dictionary and the adjustment of management vocabulary in earnings calls unanswered. Hence, I caution readers against drawing strong conclusions from these findings. However, following the results of Cao et al. (2020), this experiment indicates that managers have become more careful in their wordings. For researchers and practitioners, this may serve as another reason to move away from word lists algorithms.

6 Discussion and Conclusion

In this study, I evaluate the performance of the new FinBERT language model as financial sentiment classifier. The introduction of the finance-specific sentiment dictionary (Loughran & McDonald, 2011) can be seen as a key innovation in financial sentiment research. However, since its introduction, NLP has gone through substantial advances. So far, the financial and accounting tone literature seems hesitant to adopt new NLP models. With the recent introduction of Google's state-of-the-art BERT language model, a new opportunity to utilise machine learning in financial NLP research arises. The finance-specific version, FinBERT, applies bidirectional training of Transformers on vast amounts of financial texts. The performance of BERT-based models on various NLP tasks indicates that it has a deep sense of language.

In theory, the shortcomings of the LM Dictionary in relation to FinBERT are obvious. Dictionaries lack word context comprehension. It is well-understood that language is complicated and that words require contexts in other to be correctly interpreted. BERT is specifically trained to learn contextual relations between words in texts. One other thing to consider is the NLP arms race between investors and firms. NLP algorithms that are based on word lists can easily be fooled by simply avoiding specific terms. Managers choosing their vocabulary based on sentiment dictionaries may be problematic for future research.

I perform various experiments to test whether the theoretical superiority of FinBERT is measurable in practice. My first aim is to test the classification accuracy on labelled sentences compared to two popular sentiment dictionaries, the Harvard IV-4 Dictionary and the domain specific LM Dictionary. I find that the performance of FinBERT on two data sets with financial texts sentences is significantly better than that of two dictionary approaches. As expected, the dictionaries fall short on sentences where the context of the words is crucial. That makes FinBERT a more trustworthy tone classifier on the sentence level.

This study also examines the statistical power and the economic significance of textual tone measured by the LM Dictionary and FinBERT. I run short-window market reaction and analyst output regressions models using a sample with over 25,000 conference calls hosted by S&P 500 firms. Consistent with previous research, I show that the LM Dictionary achieves adequate results on sufficiently large data sets. That, combined with the transparency and ease-of-use that the financial dictionary offers, makes it a satisfactory financial tone model in many settings. However, researchers can significantly increase the power of linguistic tone in their test by using FinBERT. Additionally, when a small sample size is a concern, FinBERT is the better option.

I report using FinBERT results in a significantly lower probability of type II errors. Thus, the statistical power of textual tone measured by FinBERT is greater than that measured by the LM Dictionary. Further, my results show that earnings call tone constructed by FinBERT has more predictive power in all model specifications.

Additionally, I explore the potential presence of managerial tone manipulation in my sample. I find novel evidence for the hypothesis that managers avoid negative words in the LM Dictionary. In my difference-in-difference regression design, I use managers as the treated group and analysts as the control group. The results suggest that the frequency of words in the negative word lists of the LM Dictionary used by managers declined since the publication of the dictionary in 2011. This is in line with the managing-sentiment hypothesis. Future work can explore the impact of management sentiment management on financial sentiment research. It is important to know whether this phenomenon can bias future empirical work. It may serve as another argument to move away from dictionaries in tone measuring studies.

In conclusion, when the model's accuracy is more important than its transparency, FinBERT should be preferred over sentiment dictionaries. Regardless, FinBERT can be a worthwhile robustness check to alleviate concerns of measurement error in all studies that employ the LM Dictionary as their primary tone measure. As there are indications that the LM Dictionary is subject to sentiment manipulation, FinBERT may also mitigate biases that this causes, although the extent to which FinBERT itself can be manipulated is unclear.

Future research could continue the evaluation of FinBERT as textual sentiment classifier. First, one could use intraday stock returns to observe the effect of tone more directly. Second, this study can be emulated with a different sample. I use a sample of earnings call transcripts of large U.S. firms. Although my results are robust, it is unclear how well my findings generalise to samples of corporate discloses. Third, future work could pre-train FinBERT in another language to employ it on non-English financial communications. However, labelled phrases to fine-tune FinBERT on as financial text sentiment classifier are scarce in other languages.

This study is limited to only one of the many possible uses of FinBERT. One of the strengths of a FinBERT is that it is already pre-trained, and it requires relatively minimal effort to fine-tune it on a specific task. BERT has been applied successfully in text summarising, Q&A bots, and named-entity recognition. So, building on the success of FinBERT as a sentiment classification model, the path to more widespread applications is open. For instance, future research could fine-tune FinBERT for financial distress identification or stock price prediction.

References

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance*, *72*, S132–S147.

Brown, Hillegeist, S. A., & Lo, K. (2004). Conference Calls and Information Asymmetry. *Journal of Accounting and Economics*, *37*(3), 343–366.

Brown, Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … others (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2020). *How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI* (Tech. Rep.). National Bureau of Economic Research.

Chen, J. V., Nagar, V., & Schoenfeld, J. (2018). Manager-analyst Conversations in Earnings Conference Calls. *Review of Accounting Studies*, *23*(4), 1315–1354.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research*, *29*(3), 845–868.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to Forget: Continual Prediction with LSTM.

Hájek, P., & Olej, V. (2013). Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. In *International Conference on Engineering Applications of Neural Networks* (pp. 1–10).

Hassan, T. A., Hollander, S., Van Lent, L., Schwedeler, M., & Tahoun, A. (2020). *Firm-Level Exposure to Epidemic Diseases: Covid-19, SARS, and H1N1* (Tech. Rep.). National Bureau of Economic Research.

Henry, E., & Leone, A. J. (2016). Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *The Accounting Review*, *91*(1), 153–178.

Hollander, S., Pronk, M., & Roelofsen, E. (2010). Does Silence Speak? An Empirical Analy-

sis of Disclosure Choices during Conference Calls. *Journal of Accounting Research*, *48*(3), 531–563.

Huang, Teoh, S. H., & Zhang, Y. (2014). Tone Management. *The Accounting Review*, *89*(3), 1083–1113.

Huang, Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, *89*(6), 2151–2180.

Jegadeesh, N., & Wu, D. (2013). Word Power: A New Approach for Content Analysis. *Journal of financial economics*, *110*(3), 712–729.

Jung, M. J., Wong, M. F., & Zhang, X. F. (2018). Buy-Side Analysts and Earnings Conference Calls. *Journal of Accounting Research*, *56*(3), 913–952.

Kimbrough, M. D. (2005). The Effect of Conference Calls on Analyst and Market Underreaction to Earnings Announcements. *The Accounting Review*, *80*(1), 189–219.

Kothari, S. P., Shu, S., & Wysocki, P. D. (2009). Do Managers Withhold Bad News? *Journal of Accounting Research*, *47*(1), 241–276.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, *36*(4), 1234–1240.

Li. (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.

Li, Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., … Xiang, T. (2020). Enhancing BERT Representation with Context-Aware Embedding for Aspect-Based Sentiment Analysis. *IEEE Access*, *8*, 46868–46876.

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.

Luo, Y., & Zhou, L. (2020). Textual Tone in Corporate Financial Disclosures: a Survey of the Literature. *International Journal of Disclosure and Governance*, *17*(2), 101–110.

MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, *35*(1), 13–39.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *Com-*

panion Proceedings of The Web Conference 2018 (pp. 1941–1942).

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

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.

Mayew, W. J., Sethuraman, M., & Venkatachalam, M. (2020). Individual Analysts' Stock Recommendations, Earnings Forecasts, and the Informativeness of Conference Call Question and Answer Sessions. *The Accounting Review*, *95*(6), 311–337.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: from Lexicons to Transformers. *IEEE Access*, *8*, 131662–131682.

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).

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking & Finance*, *36*(4), 992–1011.

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*(3), 1193–1223.

Schumaker, R. P., & Chen, H. (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems*, *27*(2), 1–19.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645–3650).

van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1823–1832).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*,

5998–6008.

Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional Sentiment Analysis using a Regional CNN-LSTM Model. In *Proceedings of the 54Th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)* (pp. 225–230).

White, H. (1980). A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Wigglesworth, R. (2020, Dec). Robo-surveillance shifts Tone of CEO Earnings Calls. *Financial Times*. Retrieved from https://www.ft.com/content/ca086139-8a0f-4d36-a39d-409339227832

Yang, Y., UY, M. C. S., & Huang, A. (2020). Finbert: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., … Liu, T. (2019). Incorporating BERT into Neural Machine Translation. In *International conference on learning representations*.

Appendices

Definitions

Table A1: Definitions of variables.

(A) Tone variables	
$Tone^{LM}$	Tone as determined by finance-specific dictionary. [Data: Loughran and McDonald (2011)]
$Tone^{FinBERT}$	Tone as determined by FinBERT. [Pre-trained model weights: Yang et al. (2020), fine-tune data set: Huang, Teoh, and Zhang (2014)]
(C) Dependent variables	
$CAR[-1,+1]$	Cumalative abnormal return from one day before the earnings call until one day after the call. Measured by the Adjusted Market model. [Data: CRSP]
Recommendation	Change in analyst stock recommendation level. [Data: I/B/E/S]
PriceTarget	Change in analyst price target scaled by previous price target. [Data: I/B/E/S]
EPSForecast	Change in analyst earnings per share (EPS) forecast scaled by previous EPS forecast. [Data: I/B/E/S]
(C) Control variables	
Surprise	Earnings surprise scaled by the stock price. [Data: I/B/E/S & Compustat]
Size	Natural logarithm of the market capitalisation (common shares outstanding multiplied with the stock price). [Data: Compustat]
ROA	Return on Assets calculated as the ratio between income before extraordinary items and the book value of assets. [Data: Compustat]
Leverage	The ratio between long-term debt and total assets. [Data: Compustat]
Loss	Dummy variable that indicates whether there was a loss in the relevant quarter (i.e. negative net income). [Data: Compustat]
Capex	Capital expenditures scaled by total assets. [Data: Compustat]

Table A2: Sample year distribution.

	Frequency	Percentage	Cumulative		Frequency	Percentage	Cumulative
2004	1,370	5.22%	5.22%	2012	1,591	6.06%	51.86%
2005	1,451	5.53%	10.75%	2013	1,605	6.12%	57.98%
2006	1,455	5.55%	16.30%	2014	1,584	6.04%	64.01%
2007	1,497	5.71%	22.00%	2015	1,561	5.95%	69.96%
2008	1,537	5.86%	27.86%	2016	1,566	5.97%	75.93%
2009	1,550	5.91%	33.77%	2017	1,587	6.05%	81.98%
2010	1,580	6.02%	39.79%	2018	1,592	6.07%	88.05%
2011	1,575	6.00%	45.79%	2019	1,599	6.09%	94.14%
2012	1,591	6.06%	51.86%	2020	1,537	5.86%	100.00%

Confusion Matrices

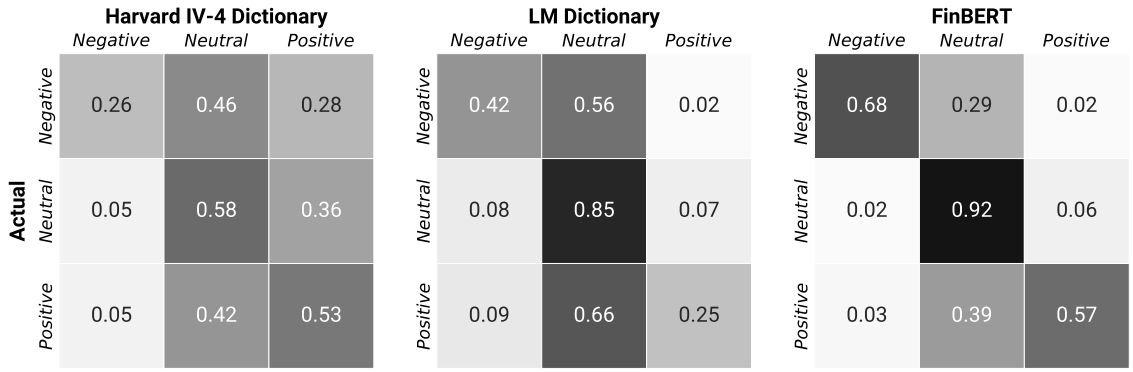

Figure A1: Financial PhraseBank (at least 50% agreement) confusion matrix. Normalized by actual labels.

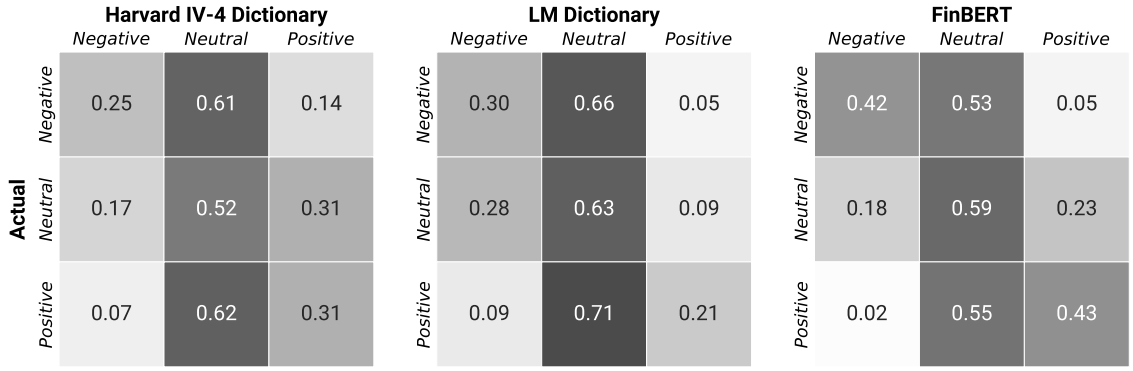

Figure A2: FiQa Posts confusion matrix. Normalized by actual labels.

LM Dictionary Word Accuracy

Table A3: Accuracy of Positive LM Dictionary phrases. This table displays the positive phrases with the highest and lowest accuracy. Requires sentences to contain precisely one word from the LM Dictionary. Only lists words with at least 5 occurrences.

		Actual Labels			
Phrase	Frequency	Positive	Neutral	Negative	Accuracy
Highest Accuracy					
improved	17	17	0	0	100.00%
boosted	7	7	0	0	100.00%
outperform	8	8	0	0	100.00%
positive	35	33	1	1	94.29%
good	33	28	4	1	84.85%
happy	6	5	1	0	83.33%
efficient	6	5	1	0	83.33%
pleased	10	8	2	0	80.00%
Lowest Accuracy					
opportunities	5	2	3	0	40.00%
innovative	6	2	4	0	33.33%
popular	6	2	4	0	33.33%
enable	14	3	11	0	21.43%
enabling	5	1	4	0	20.00%
enables	13	2	11	0	15.38%
effective	8	1	6	1	12.50%
invention	5	0	5	0	0.00%

Table A4: Accuracy of Negative LM Dictionary phrases. This table displays the negative phrases with the highest and lowest accuracy. Requires sentences to contain precisely one word from the LM Dictionary. Only lists words with at least 5 occurrences.

		Actual Labels			
Phrase	Frequency	Positive	Neutral	Negative	Accuracy
Highest Accuracy					
warning	12	0	0	12	100.00%
dropped	15	1	1	13	86.67%
lost	5	1	0	4	80.00%
weak	12	2	1	9	75.00%
layoffs	7	1	1	5	71.43%
recall	10	2	1	7	70.00%
negative	19	6	0	13	68.42%
declined	15	1	4	10	66.67%
Lowest Accuracy					
late	7	1	6	0	0.00%
divested	7	0	7	0	0.00%
disclose	12	0	12	0	0.00%
claims	5	1	4	0	0.00%
bridge	15	9	6	0	0.00%
disclosed	24	0	24	0	0.00%
challenge	5	1	4	0	0.00%
divestment	10	3	7	0	0.00%