Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Finance

Soft information in public firms' quarterly earnings calls

Isolating the effects of value-relevant information and linguistic tone in financial text

---

Otto Kopra

Master's thesis
2023

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Otto Kopra

Title of thesis Soft information in public firms' quarterly earnings calls: Isolating the effects of value-relevant information and linguistic tone in financial text

Programme Master of Science

Major Finance

Thesis advisor Prof. Samuli Knüpfer

Date 29.12.2023 **Number of pages** 86 **Language** English

Abstract

In this thesis, I study the effects that the soft information in public firms' quarterly earnings call transcripts have on stock returns. More specifically, I measure both the value-relevant information (sentiment) communicated in the earnings calls as well as the linguistic tone of the statements, by studying firms' cumulative abnormal returns following quarterly earnings calls, both on portfolio-level and with a regression analysis on earnings call -level.

My findings can be divided into methodological contributions and the uncovered financial phenomena. Firstly, I provide strong evidence in support of utilizing large language model -based solutions in extracting sentiment information from financial texts, as opposed to the dictionary-based methods that have been predominant in the financial academic literature so far. Furthermore, I show that using a tool like this together with a dictionary-based method that is designed to measure the linguistic tone of financial text makes it possible to isolate the two effects.

Leveraging these benefits of my approach, I recognize three distinct stock return phenomena related to either earnings call sentiment or tone. Firstly, I find a strong positive relation between the sentiment of an earnings call Q&A session and the cumulative abnormal returns for that firm. Moreover, unlike some of the previous studies, I find that this stock price impact is nowadays practically immediate, with this information being incorporated into stock prices in the initial reaction period. Secondly, I find evidence that the sentiment information in the calls' presentation section leads to a relatively similar immediate stock price reaction, but then to a subsequent reversal, that makes the long-term informational value in this section non-significant. Finally, I find that when controlling for the value-relevant information, a higher linguistic tone of the call's presentation section predicts lower abnormal returns, in line with the theory of strategic communication, which states that managers use of a more positive tone constitutes bad news due to their incentives to downplay negative news. I find this negative stock price reaction to be immediate only for firms that reported a negative earnings surprise, whereas firms that reported positive earnings surprise this effect takes longer to be incorporated into share prices.

Keywords earnings call, sentiment analysis, linguistic tone

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Otto Kopra

Työn nimi Pehmeä tieto julkisten yritysten kvartaalitulosjulkistusten puhelinkonferensseissa: Arvorelevantin informaation ja kielellisen sävyn vaikutusten eriyttäminen rahoitusalan tekstissä

Koulutusohjelma Kauppatieteiden maisteri

Pääaine Rahoitus

Työn ohjaaja Prof. Samuli Knüpfer

Päivämäärä 29.12.2023 **Sivumäärä** 86 **Kieli** englanti

Tiivistelmä

Tässä opinnäytetyössäni tutkin julkisten yritysten kvartaalitulosjulkistusten yhteydessä pidettävien puhelinkonferenssien pehmeän tiedon vaikutuksia osakkeiden tuottoihin. Mittaan erityisesti sekä puheluissa välitettyä arvorelevanttia tietoa (”sentiment”), että puheen kielellistä sävyä (”tone”), tutkimalla yritysten kumulatiivisia ylisuuria tuottoja puhelukonferenssien jälkeen sekä portfoliotasolla että regressioanalyysin kautta puhelinkonferenssitasolla.

Työni tulokset voidaan jakaa metodologisiin kontribuutioihin sekä havaittuihin taloudellisiin ilmiöihin. Ensinnäkin esitän vahvaa näyttöä suurten kielimallipohjaisten ratkaisujen hyödyntämiselle sentimentti-tiedon mittaamiselle rahoitusalan teksteissä, toisin kuin taloustieteellisessä kirjallisuudessa tähän asti vallinneet sanakirjapohjaiset menetelmät. Lisäksi näytän, että käyttämällä tällaista työkalua yhdessä sanakirjapohjaisen menetelmän kanssa, on mahdollista mitata paremmin molempia ilmiöitä – sekä arvorelevanttia informaatiota, että kielellistä sävyä.

Hyödyntämällä näitä lähestymistapani hyötyjä havaitsen kolme erillistä osaketuottoilmiötä, jotka liittyvät joko konferenssipuhelun sentimenttiin tai kielelliseen sävyyn. Ensinnäkin löydän vahvan positiivisen yhteyden tulospuhelun Q&A-istunnon sentimentin ja ko. yrityksen kumulatiivisen ylisuuren tuoton välillä. Toisin kuin monissa aiemmissa tutkimuksissa, havaitsen tämän reaktion osakkeen tuotoissa tapahtuvan nykyään käytännössä välittömästi. Toiseksi löydän näyttöä sille, että puheluiden alkuosion sentimentti johtaa suhteellisen samanlaiseen välittömään tuottoreaktioon, mutta tämän jälkeen päinvastaiseen käänteeseen, mikä tekee informaatioarvosta tässä osiossa pitkällä aikavälillä merkityksettömän. Lopuksi osoitan, että kun puhelun arvorelevanttia tietoa kontrolloidaan regressiossa, puhelun alkuosion korkeampi kielellinen sävy ennustaa negatiivista ylisuurta tuottoa. Tämä on linjassa strategisen viestinnän teorian kanssa, jonka mukaan johdon positiivisen sävyn käyttö on ymmärrettävissä huonoksi merkiksi, sillä johdolla on kannustin vähätellä huonoja uutisia. Tämä negatiivinen osakekurssireaktio on välitön vain yrityksille, jotka ilmoittavat negatiivisesta tulosyllätyksestä, kun taas yritykset, jotka ilmoittivat positiivisesta tulosyllätyksestä tämän vaikutuksen sisältyminen osakkeiden hintoihin kestää kauemmin.

Avainsanat konferenssipuhelu, sentimenttianalyysi, kielellinen sävy

Table of contents

1 Introduction

While acknowledging the risk of oversimplification, it is evident that academic research in finance has predominantly centred around quantitative analysis of numerical data. The financial markets have been studied thoroughly to understand how numbers related to for example firm's characteristics – such as size or book-to-market equity – or events – such as earnings surprise – affect another figure, like the firm's share prices or stock returns. While a lot of the information available to financial markets about a company is indeed numerical, this "hard" information coexists with its counterpart: "soft" information. This soft information is any type of non-quantifiable or subjective data about a company. In this thesis, I study a specific type of soft information: the textual tone and sentiment in firms' quarterly earnings calls. Accompanying the quarterly earnings announcements, these earnings calls earnings calls serve as a crucial information channel for investors and analysts seeking insights into a company's present state and future prospects. During these calls, analysts strive to delve deeper into the inner workings of firms beyond the disclosed financials, while management endeavours to convey this information through a credible narrative.

In this thesis, I first study whether a new type of sentiment analysis tool utilising deep learning, a Large Language Model (LLM) -based FinancialBERT can be used to provide a better measurement for the sentiment of financial text when compared to previous methods. I then explore whether this tool makes it possible to decouple the "abnormal" linguistic tone of a text from the value-relevant information (sentiment) in financial textual data. Finally, I study how these two aspects of soft information in quarterly earnings calls affect public firms' stocks returns and how this information is incorporated into share prices.

The purpose of this thesis is (i) to demonstrate the utility of using large language models in content analysis of financial textual data, (ii) to propose a new distinction between often interchangeably used terms of "tone" and "sentiment" and (iii) advance our understanding of the financial market's reaction to the soft information in earnings calls.

For example, Price et al. (2012) have previously studied the textual tone and its informativeness in earnings calls between 2004 and 2007. Their main finding was that the call's tone has incremental information value, and that it predicts abnormal stock returns and abnormal trading volume during the earnings announcement window. Furthermore, they find that this information is not fully incorporated into share prices immediately, but rather do so slowly over the post-earnings announcement drift (PEAD) period, during the 60 trading days following the earnings call. On the other hand, Blau et al.'s (2015) study a hypothesis that inflated (i.e. overly positive) talk is a form of managers' strategic communication and should be considered bad news. They find that sophisticated investors target firms with both high earnings

surprise as well as high abnormal tone, and that in the presence of both these factors together increases short sellers' return predictability. Both Price et al. (2012) and Blau et al. (2015), together with the vast majority of existing literature, use a so-called dictionary-based approach, which relies on counting the number of predetermined positive and negative words in a given text, and using those word counts to calculate a proxy for the overall positivity/negativity of the text. For Price et al. (2012), this dictionary is Henry's (2008) wordlist, whereas Blau et al. (2015) use the Loughran & McDonald (2011) wordlist.

Notwithstanding the expanding body of literature, the field of financial sentiment analysis exhibits certain gaps, the first one being a methodological one: Unlike studies utilising computational linguistics tools in many other disciplines, the finance and accounting literature is still strongly dominated by rudimentary bag-of-words -approaches, which has drawn criticism from some (e.g. El-Haj et al. 2019). In many other academic fields, computational linguistics analysis is increasingly often done utilising more sophisticated methods, such as deep learning and other machine learning solutions, which at their best exhibit superior capabilities when compared to alternative methods. Furthermore, there seems to be a discernible lack of enthusiasm among academics to more accurately discuss or define their interpretation of the tone or sentiment they aim to quantify. An early work by Tetlock (2007) discusses the two alternative explanations for the existence of a high sentiment: the information theory posits that the high sentiment proxies real value-relevant information about the stock market, whereas the sentiment theory asserts that the sentiment measures individual investors' behavioural attitudes towards the stock market. However, most studies are ambiguous about what exactly they mean by tone or sentiment. Lastly, the findings presented by Price et al. (2012) during a sample period shortly after the public accessibility of earnings calls have not undergone much examination more recently. During the past two decades of public availability of these calls, financial markets have likely developed a better understanding about their informativeness. Concurrently, within the same timeframe, certain other market inefficiencies, such as the post-earnings announcement drift, have been demonstrated to have largely disappeared (Martineau 2021), making this phenomenon that much more interesting to study.

In this thesis, I use earnings calls from firms of three major US stock exchanges during a period ranging from 2008 to 2021. For each earnings call, I determine a "tone" score utilising Henry's (2008) wordlist, as well as a "sentiment" score utilising Hazourli's (2022b) FinancialBERT deep learning model. I then analyse how these two scores explain the stock's cumulative abnormal returns over two reaction time periods: The initial reaction period, from day -1 to day 1 around the earnings call, and the extended reaction period, from day 2 to day 60 after the earnings call. Additionally, I also conduct some tests for the full (combined) reaction period, from days -1 to day 60. I

initiate this analysis by first dividing the sample into portfolios based on the tone and sentiment measures. I then conduct tests for the differences of means and medians as well as visual assessment of the reaction periods' returns. After this, I regress the cumulative abnormal returns on the tone and sentiment measures both for the full sample as well as for a set of cross-sections of the sample.

The primary findings of this thesis are as follows. Firstly, I find strong support in favour of using LLM-based FinancialBERT instead of Henry's tone to assess the soft information content of financial texts. Looking at the stock returns in the initial reaction period, the use of FinancialBERT instead of Henry's tone measure improves the regression model's explanatory power by roughly 30%. In economic terms, a single standard deviation change in the tone/sentiment measure for the Q&A section of an earnings call is associated with a 1.02% change in the cumulative abnormal returns when using FinancialBERT, compared to a 0.70% change observed with the dictionary approach. Secondly, I find that the when controlling for value-relevant information with FinancialBERT sentiment, Henry's tone measure can be used to capture the abnormal linguistic tone component of the text. My findings show a negative connection between this tone measure for the presentation section of a call and the stock's cumulative abnormal returns. A single standard deviation change in the abnormal tone is associated with a -0.57% change in cumulative abnormal returns during the initial reaction period. Moreover, supplementary analysis indicates that how this phenomenon is incorporated into share prices varies between different firms: The stock market appears to factor in this information in the initial reaction period only for firms with a negative earnings surprise. Conversely, for firms with a positive earnings surprise, this information is incorporated into share prices more slowly over the extended reaction period. Finally, my results suggest that the overall stock price "drift" on soft information documented by Price et al. (2012) does not exist anymore, but the value-relevant soft information in earnings calls' Q&A section is incorporated into stock prices fully during the 3-day reaction period. To the contrary, the results suggest that there is instead an immediate overreaction to the sentiment information in the calls' presentation section, with a subsequent reversal over the extended period.

Academically, my results provide several lessons regarding how soft information in financial texts should be understood and studied. Firstly, it is important for researchers to recognize that when studying sentiment or linguistic tone, it would be useful to distinguish between the different phenomena and use methodologies that allow for the separation of these effects. Secondly, my findings support the idea that academics should embrace newer and more sophisticated methodologies when conducting their research agendas in computational linguistics. For a more practical significance, my results offer several insights into informational efficiency of the financial markets, providing potentially useful insights to managers, investors, and financial

market regulators on how soft information in earnings calls is conveyed, understood, and reacted to by the market participants.

The rest of the paper is structured as follows: Chapter 2 provides the overview of the existing literature, including a more in-depth explanation of the different textual content analysis methodologies used in the finance and accounting field. In chapter 3 I explain the motivation, research questions, hypotheses as well as the main contributions of this study. Chapter 4 describes the data and methodologies used in this study. Chapter 5 presents the main empirical analysis and discussion about its results. In chapter 6 I introduce some additional analysis to provide further insights into the discovered phenomena, and chapter 7 concludes this thesis.

2 Literature review

In this chapter, I will go through existing literature that is relevant to this study. In section 2.1 I will summarize the literature on the post-earnings announcement period stock returns. Section 2.2 presents an overview of the current literature on textual content analysis in finance. Section 2.3 summarizes the literature in financial studies related to measuring linguistic tone or sentiment. In section 2.4 I discuss and present literature on the theoretical definitions for tone and sentiment that existing studies have used. In section 2.5 I provide an overview of literature related to traditional methodologies for financial sentiment analysis, and finally in section 2.6 I discuss some of the newer avenues of research and more modern methods for financial sentiment analysis.

2.1 Literature on the post-earnings announcement period

In 1968, Ball & Brown first documented their finding that following firms' earnings announcements, stock prices would "drift" in the same direction as the unexpected earnings "surprise". This idea, seemingly at odds with the efficient market hypothesis indicated that it took some time for financial markets to fully incorporate publicly available information into share prices. Subsequently, the post-earnings announcement drift (PEAD) has become a popular topic in financial academic literature. Some studies attribute PEAD to phenomena that can arise from informationally efficient markets, such as transaction costs (e.g. Bhushan 1994), arbitrage risks (e.g. Mendenhall 2004), or illiquidity (e.g. Chordia et al. 2009; Sadka 2006). Other studies, however, explain the PEAD with market's failures to understand new information; research by both DellaVigna & Polle (2009) and Hirshleifer et al. (2009) point to limited investor attention as a reason for PEAD. Abarbanell & Bernard (1992) find evidence that even professional analysts underreact to recent earnings, hinting that slow learning and information processing might be behind PEAD. Bernard & Thomas's (1989) results showed PEAD to be more consistent with a delay in the market response to earnings reports than a risk-premium based explanation.

While PEAD has been shown to be surprisingly persistent, it seems that PEAD has largely vanished during the past two decades. Martineau (2021) documents exactly this disappearance, showing the drift diminishing during the traditional 60-day period after the earnings announcement. Studying market inefficiencies more generally, Chordia et al. (2014) report a decline in the profits for anomaly-based trading strategies, and Mclean & Pontiff (2016) find a post-publication attenuation of academically published anomalies. All of this suggests that markets have developed to become informationally more efficient.

2.2 Literature on textual content analysis in finance

While the traditional post-earnings announcement drift might have disappeared, that doesn't mean that all the information in earnings announcements gets incorporated into stock prices immediately. After all, as Martineau (2021, forthcoming) states, the surprise in the earnings measure (most often earnings per share), is nothing more than "a noisy proxy of the 'hard' information content embedded in earnings announcements". Martineau notes that the so-called "soft" information elements, such as earnings conference calls, may require more time for market participants to incorporate into share prices. Indeed, there has been increased attention among academics to try to understand different types of soft information. The different data that academics have studied include a range of firm disclosures, such as annual reports (e.g. Loughran & McDonald 2011 and 2014; Li 2008; D'Augusta & DeAngelis 2020), earnings press releases (e.g. Davis et al. 2012; Henry 2008; Tan et al. 2014; X. Huang et al. 2014), as well as alternative sources of soft information, such as news articles (e.g. Tetlock 2007; Tetlock et al. 2008; Garcia 2013) or social media posts (e.g. Chen et al. 2014).

While studying these types of textual datasets is something that can be done manually, a more common approach (especially with larger datasets) is to utilise computational linguistics, which refers to the analysis of language and speech using techniques of computer science. Of academic papers in accounting and finance disciplines that leverage computational linguistics, the most studied texts are annual and quarterly reports followed by conference calls (El-Haj et al. 2019). Firms' quarterly earnings calls are often regarded as an appealing resource for a couple of reasons: First, the earnings calls serve as an important source of information about firms to market participants, especially since the enactment of Regulation FD in 2000, which made the calls public in (Frankel et al. 1999; Mohanram & Sunder 2006; Irani 2004). Another reason relates to for example Davis & Tama-Sheet's (2012) findings that managers use more optimistic tone in earnings press releases when it is beneficial to them. In addition to the pre-prepared management presentation, earnings calls typically encompass a Q&A session, providing analysts with an opportunity to pose questions to the firm's management. This Q&A session is understood to be less susceptible to the use of inflated language compared to other forms of investor communications, as the management's ability to control the discourse is limited. For example, Price et al. (2012) and Blau et al. (2015) emphasize the dynamic nature of the Q&A portion of the call, where analysts can challenge the management's statements. Indeed, Brockman et al. (2015) find evidence that managers' tone tends to be more optimistic than analysts' tone during earnings calls.

There are several different aspects of soft information that can be extracted from textual data. Academics have previously studied for example the information in forward-looking statements (e.g. Li 2010; Muslu et al. 2014;

Athanasakou & Hussainey 2014; Schleicher & Walker 2010; Baginski et al. 2004), riskiness or uncertainty (e.g. Kravet & Muslu 2013; Hope et al. 2016; Hassan et al. 2019; Hanley & Hoberg 2012; Campbell et al. 2013), as well as readability or "fogginess" of the disclosure (e.g. Guay et al 2016; Tan et al. 2015; Rennekamp 2015; Loughran & McDonald 2014; Lo et al. 2017).

2.3 Literature on linguistic tone and sentiment analysis

By far the most popular feature for academics' use of computational linguistics methods has however been the "tone" or "sentiment" of text (El-Haj et al. 2019). Kearney & Liu (2014) define that "tone" is the positivity or negativity of the source material, whereas "textual sentiment may also include affects other than positivity-negativity, such as strong-weak, and active-passive". However, in the academic literature these two terms have been used mostly interchangeably to refer to the overall level of positivity or negativity of a text.

Sentiment analysis has been done for a range of different textual data to reveal different characteristics of market or firm-level data. One popular area is to connect the sentiment of the texts to firms' stock returns or other stock data, such as trading volume or volatility. These studies use sentiment analysis to sources including annual or quarterly reports (e.g. Li 2010; Loughran & McDonald 2011), earnings press releases (e.g. Henry 2008; Demers & Vega 2008), news articles (e.g. Tetlock 2007; Tetlock et al. 2008), analyst reports (e.g. Engelberg et al. 2012; Twedt & Rees 2012) internet board postings (e.g. Antweiler & Frank 2004; Das & Chen 2007), and other sources (e.g. Ferris et al. 2013 study tone in IPO prospectuses; Kothari et al. 2009 and Jiang et al. 2019 use a mixed set of sources; Baginski et al. 2018 study management forecasts). Nevertheless, as previously noted, earnings calls can offer a distinctly less biased perspective on a firm.

Price et al. (2012) study the incremental informativeness of textual tone in quarterly earnings calls, relative to the earnings surprise between 2004 and 2007. They find that the tone of both the call's presentation part and the Q&A part have incremental information value, predicting abnormal stock returns and abnormal trading volume during the announcement window. Moreover, the authors find that the information in the call tone is not fully incorporated into share prices immediately, but rather do so slowly over the PEAD period, during the following 60 trading days from the call. They note that this is consistent with e.g. Engelberg (2008) and Demers & Vega (2008), in that market participants find it more challenging to comprehend qualitative information, and as a result, this type of information is fully incorporated into market prices only after a slightly longer period of time. Price et al.'s (2012) findings that equity markets react to linguistic tone are corroborated by for example Doran et al. (2012) who study call tone in REITs, and Borochin et al. (2018), who find connection between positive call tone and options-based measures for firm value uncertainty. Yamamoto et al. (2022)

create a long-short trading strategy based on call tones that yields 7.07% annualized returns. Correspondingly, Fu et al. (2021) find evidence that low tone in year-end earnings call predicts higher stock price crash risk in the following year, and Fei et al. (2023) find that suppliers give more trade credit to customers whose managers' have a higher tone in earnings calls.

However, when it comes to abnormal component or extremities of tone, the findings are somewhat more mixed. Larcker et al. (2012) find that deceptive CEOs use much more extreme positive tone than others. Lee (2016) finds that in the presence of bad news, managers are more likely to fall back on scripted answers, speaking less spontaneously. On the other hand, Bochkay et al. (2020) find that managers' use of "extreme" words is associated with increased trading volume and a stronger positive share price reaction, although these effects are stronger for firms with weaker information environment.

Blau et al. (2015) base their hypotheses around ideas from Kartik et al. (2007), who showed that under their model of strategic communication, managerial tone is overly optimistic, which is deceiving naïve followers. Under the authors' hypothesis, inflated talk should be considered bad news. The authors measure this "inflated talk", or abnormal tone as the difference between the management presentation and the Q&A sections of the calls and find that sophisticated investors (short sellers) target firms with both high earnings surprise as well as high abnormal tone. Furthermore, they find that in the presence of both high earnings surprise and abnormal tone, short sellers' return predictability is increased. These results suggest that sophisticated investors are more able to accurately assess inflated talk as the bad news that it is, but also that managers cannot support a prolonged overvaluation of their firms with high tone. While the authors do not show a direct negative link between abnormal tone and stock returns, the results nevertheless suggest that an abnormally high earnings call tone should indeed be considered as a bad sign for firm value.

These questions about abnormal tone have also been studied with other forms of textual data. X. Huang et al. (2014) study the impact of abnormal tone in earnings press releases. The authors define "abnormal tone" as the part of tone that is not explainable by a combination of a set of quantitative measures, such as total assets, past stock returns, book-to-market equity, return volatility and other factors. They find that a higher abnormal tone has an immediate positive stock return reaction, but a subsequent negative effect in the next 120 days, showing the negative valuation implications of a high abnormal tone.

Following the same methodology for determining abnormal tone, Baginski et al. (2018) study publicly released management forecasts. Parallel to the findings of Blau et al. (2015), the authors find that different investors seem to assess abnormal tone in varying ways. Their results suggest that in the presence of higher abnormal tone, small (large) investors engage

relatively more in buying (selling) activity. the paper also shows abnormal tone having a negative impact on stock returns, both during the immediate 3-day reaction window, as well as during an extended 120-day drift period after the disclosure.

Naturally, measuring abnormal tone by defining "abnormality" as anything not explainable by a set of quantitative measures ignores the fact that there might be incremental value-relevant information within the textual data. Hennig et al. (2023) study tone in earnings calls but also include linguistic features proxying for managers' and their statements' credibility in their analysis. Not unlike the above-mentioned studies focusing on abnormal tone, the authors find an immediate positive reaction to higher linguistic tone, and a subsequent negative reaction during the extended reaction period. However, this reversal was found to be smaller for statements with higher credibility measures.

So, what conclusions can be derived from all this? Firstly, a linguistic tone as a single measure is incomplete in describing the different ways that the managers' use communication conveys information to financial markets. While a higher tone is frequently associated with increased stock returns, an abnormally high tone may yield the opposite effect. Indeed, there seems to be two different phenomena at play. Secondly, the existing ways to measure the "abnormal" component of tone are far from perfect. The most used approach is the one by X. Huang et al. (2014), which assumes that any part of tone not explainable by firm fundamentals and specific set of quantitative measures is "abnormal" in nature. Whether this abnormal tone serves as a proxy for value-relevant information or constitutes biased discourse is not inherently determined by the measure itself. Rather, such distinctions might have to be construed post hoc by examining factors that may explain the observed results. Thirdly, it is well established now that the incorporation of soft information into share prices has been happening only partly immediately, and that the stock prices have accounted for all the tone information only after a considerable delay, which is evident in the different drifts and reversals observed by academics. Finally, it is safe to say that it is not know too well how abnormal tone in quarterly earnings calls affects firm valuations. Blau et al. (2015) do study this phenomenon and provide a relatively innovative approach of using the tone of the earnings calls' Q&A section as the baseline and calculating the abnormal component of the presentation section's tone. However, the authors do not show a direct connection between their abnormal tone measure and the subsequent stock returns.

Considering all these factors, there seems to be a gap in the literature in understanding the different phenomena that might be affecting firm's stock returns simultaneously. Much of the research on abnormal tone does not accommodate for the fact that managers may also convey factual and value-relevant information in their disclosures. A more comprehensive approach would measure the linguistic tone of communication while also accounting

for value-relevant information within the disclosure. This is also hinted at by Hennig et al.'s (2023) conclusion, that "…results provide evidence that investors benefit from considering credibility signals from the simultaneously perceived soft information when reacting to tone in conference calls.".

2.4 On the nature of tone and sentiment in financial literature

With all this research on textual tone or sentiment in the span of the past two decades or so, one would be forgiven for thinking that these two terms had a clearly stated and universally agreed-upon definition. Unfortunately – although perhaps not surprisingly – this is not the case.

Algaba et al. (2020) talk about sentiment on a general level and propose a generic definition: "Sentiment is the disposition of an entity toward an entity, expressed via a certain medium", whereas Zhou (2018) defines sentiment mathematically as the difference between observed or expected characteristic of an asset and the same characteristic implied by a benchmark model. Zhou notes that most proxies for investor sentiment are only measures of the investors' beliefs without a benchmark, which does not help in understanding if the sentiment (and thus the beliefs) are rational or irrational.

Tetlock (2007) provides an early study that touches on the different possible explanations for sentiment and provides two different views on it: sentiment theory and information theory. When studying investor sentiment for the aggregate market with Wall Street Journal articles, Tetlock (2007) defines sentiment as level of beliefs held by "noise traders" (who hold random beliefs about future dividends) relative to those held by "rational arbitrageurs" (who hold Bayesian beliefs). If noise traders' beliefs are below Bayesian traders', Tetlock deems the noise investors "pessimistic". When Tetlock measures the news media sentiment, he tests whether high media pessimism is associated with low investor sentiment. If "sentiment theory" was true, media pessimism would either forecast negative investor sentiment, or alternatively reflect the already existing investor sentiment. In addition to the sentiment theory, Tetlock also discusses an alternative explanation for a possible connection between media pessimism and stock returns: the "information theory". This theory is based on the possibility that media pessimism acts as a proxy for actual negative information about equities fundamental values, which might or might not yet be incorporated into share prices. Ultimately, Tetlock's findings support the sentiment theory's explanation of media pessimism forecasting market sentiment, leading to downward pressure in the short term, but a reversal in the longer term. He furthermore states that the information theory interpretation that media pessimism would act as a proxy for actual information that is yet to be incorporated into share prices "receives very little support from data". However, when Tetlock et al.

(2008) conduct another study on media sentiment, only this time on a firm-level utilizing firm-specific news stories, the authors appear to assume the presence of value-relevant information in the news stories, stating that "linguistic communication is a potentially important source of information about firms' fundamental values". The authors nevertheless find evidence supporting that idea. They first show that the sentiment information in news articles does indeed include value-relevant information, that is incorporated into stock prices only after a slight delay. Furthermore, they show that this sentiment can also be used to predict both firm earnings and stock returns. Thus, there is no clear judgement in favour of either information or sentiment theory in these studies.

Henry (2008) conducts an event study, researching the earnings press releases' tones and their impact on stock returns. As this paper uses press releases, which are released by the firms' management, her theory for the expected connection between the tone and stock returns is a behavioural one, based on the prospect theory by Tversky & Kahneman (1981, 1986). Henry predicts that her measure of tone is related to the communicational framing of financial performance, with a more positive tone leading to higher returns. The author's results do indeed show a positive connection between tone and abnormal stock returns, at least up to a point. However, she defines "tone" in her study as "a function of both content and word choice", acknowledging the limitation that press releases "exhibit a potential duality of purpose: information and promotion.", and that the quantitative dictionary method used in the study is not able to understand text on a more complex level. This underscores the reality that, despite conducting the study with a clearly defined meaning for the word "tone", and utilising a purpose-built word list, the methodology's limitations still impede the certainty of the conclusions.

Henry also notes that some of the other studies in the field do not give any formal definition for their understanding of the word "tone" but are rather just treating the concept as a generally understood positive or negative nature of communication. In many studies in accounting and finance literature that use tone or sentiment in some way, same problem persists. Indeed, the indifference toward questioning the actual meaning of the terms "sentiment" and "tone" in textual analysis might not be so surprising, as it is likely influenced, at least in part, by the limitations inherent in our methodologies. There has not really been an elegant way to isolate the value-relevant information from the more sentimental linguistic tone of a text. Most studies, especially focusing on earnings calls, control for the hard numerical information, such as the earnings surprise, which does not help with controlling with the value-relevant information in the soft data that goes beyond these numbers. It is true that some studies have combined tone measures with other measures for characteristics like deceptiveness (Larcker et al. 2012) or credibility (Hennig et al. 2023) or have created some measure for "abnormal" or "residual" tone (e.g. Blau et al. 2015; Baginski et al. 2018). However, as

mentioned, many of these approaches have issues of their own, as they might easily get overly complicated, and still ultimately fail in effectively isolating the desired effect. Hence, how to distinguish between "what" is being said and "how" it is being said remains a mostly unanswered question even today.

2.5 Traditional methods in financial sentiment analysis

Kearney & Liu (2014) categorize the prevalent content analysis methods in textual sentiment analysis literature into two groups: dictionary-based approaches and machine learning. Of these, the dictionary-based approach is overwhelmingly the most utilised. The dictionary approach is sometimes also referred to as the bag-of-word approach. In this approach, the researcher has one or more wordlists ("dictionaries"). The document is then divided into individual words, and each instance of a word found in the dictionary is counted. Most often, this has meant having a word list of positive words and negative words and counting the number in each category to determine the overall tone of the text. The "bag-of-words" characterization of this method refers to the fact that all the words of the document are assessed individually, and the approach does not consider the sentences in which they appear or any other context in which they are written.

The earliest works in the literature often used a general English dictionary, such as the Harvard IV-4 Psychosocial list or word lists from a textual analysis program DICTION (Kearney & Liu 2014). A widely acknowledged (e.g. Henry & Leone 2016; Loughran & McDonald 2011) problem with these types of word lists is that words can have very different meanings or connotations depending on the discipline. Some words that are neutral in financial context, such as "liability" and "tax" are found in the list of negative words in the Harvard IV-4 dictionary (Kearney & Liu 2014).

That is why today dictionary-based studies most tend to use domain-specific dictionaries. In the field of accounting and finance, the two most used such dictionaries are Loughran & McDonald's (2011) list, and Henry's (2008) list. These dictionaries have been found to outperform the general dictionaries in assessing the tone of financial texts (e.g. Price et al. 2012; Henry & Leone 2016). These dictionary methods have been sometimes refined with additional adjustments. While most studies use proportional weights, term weighing has been used to adjust how much significance is given to each word in the document (Kearney & Liu 2014). The weights might be decided for example based on the frequency of the given word in the document (e.g. Loughran & McDonald 2011), or the known prior market reactions to each word (Jegadeesh & Wu 2013). Some other studies (e.g. Brau et al. 2016; Borochin et al. 2018) make a different tweak to the dictionary-approach by adjusting for negation, by not counting words preceded by a negation word ("not", "no", "never", etc.).

The alternative approach to dictionary-based methods is machine learning -based solutions. Kearey & Liu (2014) list some of the used methods, such as Naïve Bayesian algorithm, k-nearest neighbour, tf-idf, and probabilistic indexing. Despite the relative simplicity of these algorithms, they are still often much more cumbersome to use for researchers than dictionary-based methods. The machine learning algorithms listed above require the process of manually annotating part of the corpus as the training data. The actual required size for this training set depends on factors such as desired model accuracy, the used corpus, as well as the classification task; For a naïve Bayesian algorithm, Antweiler & Frank (2004) annotate 1,000 internet board messages' tone for their use, while Li (2010) annotates an impressive 30,000 sentences manually into different categories of tone and topic for the algorithm's training set.

Ignoring the ease of deployment, what do we know about how these machine learning methods stack up against dictionary methods? A lot depends on the actual model, but A. H. Huang et al. (2014) report that their naïve Bayesian algorithm, trained with 10,000 manually annotated sentences, reaches an accuracy of around 80%, compared to general English dictionary-methods' accuracy of around 50%. Li's (2010) results suggest that dictionary-based methods might not work very well for financial statement domain, with a naive Bayesian method reaching a better result, with a 67% accuracy for tone measurement. However, as mentioned, this far from perfect accuracy level was achieved only after manually annotating 30,000 sentences. This need for a large amount of manual work, combined with sometimes questionable level of provided improvement, might be why machine learning -based approaches are still not too often used by researchers. Of the already discussed papers on earnings calls tone (including six papers published in the past four years: Bochkay et al. 2020; De Amicis et al. 2021, Fu et al. 2021; Yamamoto et al. 2022; Fei et al. 2023; and Hennig et al. 2023) all use a dictionary approach to conduct their sentiment analysis, with Loughran & McDonald's wordlist being the most common one.

2.6 Advanced methods in financial sentiment analysis

While progressing into finance-specific dictionaries or deploying different types of machine learning algorithms does seem to improve the quality of research, the benefits are still limited. Even with these more suitable wordlists, the bag-of-words -method is unavoidably rudimentary. A key problem is that these methods treat all words independently, without any information on their conjugation, connection to other words or the context in which they are said, which shows in the relatively modest classification accuracies of these methods.

In their assessment of the current state of accounting and finance research that implements computational linguistics methods, El-Haj et al.

(2019) present strong criticism against the most used methods. The authors conclude that the "mainstream … research appears to be behind the curve in terms of CL sophistication generally, and word sense disambiguation in particular", or as they put it even more bluntly: "Promoting research agendas where the focus or approach continues to rely on bag-of-words technologies such as readability or and basic keyword content analysis, will at best represent a missed opportunity and at worst yield a body of work whose credibility and relevance is questioned by future generations." Indeed, when academics conclude their research by calling for the development of "better" or "more accurate" wordlists (e.g. Price et al. 2012; Kearney & Liu 2014) for future research purposes, they seem to inadvertently put the specific easy-to-use – but very flawed – methodology ahead of the actual need, which is just generally better tools to analyse content in financial texts.

El-Haj et al. (2019) note that one of the limitations to the development of CL techniques has been a lack of recognized corpora. The authors state that unlike many other disciplines, accounting and finance research has largely ignored this pursuit. This puts limitations on both replicability of existing studies, as well as conducting new research. For example, Li (2010) never published his manually annotated collection of 30,000 sentences. Specifically for machine learning -based methodologies, other disciplines have benefited greatly from an openly published datasets, and especially ones that have been manually annotated. This has allowed researchers to develop and test different approaches and algorithms, as well as compare their performance with the work form other researchers. For example, Deng et al. (2009) published ImageNet, a collection of tens of millions of annotated images, one of the best-known in its category that has since been used to train and improve image classification solutions. In 2012, Krizhevsky et al. (2017) presented their method for image classification task with ImageNet collection, reaching an accuracy of 62.5%, a clear ca. 10 percentage point improvement to previous methods. Currently, the best models are already scoring at above 90% accuracies (e.g. Wortsman et al. 2022). This kind of advancement is not exclusive to general computer science or mathematics fields. For medical field, datasets are numerous: MIMIC-III is an openly published database of de-identified critical care patient records (Johnson et al. 2016) that has been used to train models on for example mortality or length-of-stay -prediction, whereas ChestX-ray14 dataset introduced by Wang et al. (2017) is a collection of over 100,000 X-ray images annotated with disease labels. For understanding textual data there are plentiful datasets, such as QNLI, a question-answering dataset introduced by Wang et al. (2019). For sentiment analysis tasks, there are datasets ranging from IMDb Movie Reviews (Maas et al. 2011) to MPQA Opinion Corpus (Wiebe et al. 2005), a dataset of news articles annotated with different opinions and sentiments. While the development for publicly available datasets for machine learnings algorithms' development has been much slower in the accounting and finance disciplines, it has

not been completely non-existent. In 2014, Malo et al. published their research on accommodating the phrase-structure level information and domain-specific financial language for sentiment analysis. Along with this research, they published the "Financial Phrase-Bank" dataset. This is a collection of around 5,000 sentences extracted from financial news articles and press releases, each one tagged as either "positive", "negative" or "neutral" by a human annotator with a business education background.

In the recent past, general machine learning research has taken some major steps forward. The developments brought by technological advancements in deep learning has allowed great leaps forward (Bengio et al. 2021). Specifically, the emergence of foundational models has allowed for a huge leap forward in textual content analysis. Foundational models, such as Large Language Models (LLMs) in the context of NLP, are deep neural models that have been first "pre-trained" with a very large amount of data. It has been found that after this pre-training, these models can then be fine-tuned for specific tasks, at a relatively small cost, and still achieve high accuracy, even when the training set for the final task is relatively small (e.g. Devlin et al. 2019; Brown et al. 2020; Sun et al. 2019). One such widely used LLM is BERT (Devlin et al. 2019) introduced by Google.

As BERT has already been used as a basis for NLP tasks across the scientific fields, so too has BERT models trained specifically for financial applications emerged. In just the few past years, Araci (2019), Yang et al. (2020), Liu et al. (2021), Huang et al. (2023) and Hazourli (2022b) all presented their own versions of "FinBERT", a BERT model trained further with financial data (although differing from the others, Hazourli calls his model "FinancialBERT"). The advantage that these models have over simpler models is that the base BERT model has already been trained to understand general language with data such as BooksCorpus and the entire English-language Wikipedia, resulting in learning dataset of 3.3 billion words (Devlin et al. 2019). In Hazourli's (2022b) FinancialBERT, a BERT model initialized with these initial parameters is further pre-trained with a large amount of finance-specific texts, including a collection of news articles from Thomson Reuters and Bloomberg, corporate reports (annual and quarterly reports) as well as earnings call transcripts. All of these sources bring an additional 3.39 billion words of pre-training data to FinancialBERT. Only after this learning of both from general English sources and financial text sources, is the model then "fine-tuned", trained for the specific ultimate task, in this case sentiment analysis using the Financial PhraseBank (Hazourli 2022b). Ultimately, Hazourli's FinancialBERT seems to outperform the other models, achieving an accuracy of 99% in classifying the sentiments of the Financial PhraseBank. Other versions of FinBERT use a roughly similar training pipeline, with some differences in training data and results. For example, Liu et al.'s (2021) version is pretrained on financial news, as well as Yahoo Finance articles and finance-related posts from Reddit, resulting in a sentiment classification

accuracy of 0.94. The other models (Araci 2019; Yang et al. 2020; Huang et al. 2023) reach sentiment classification accuracies between 0.86 and 0.88, which provide marginal improvements over the base BERT model, which Hazourli (2022b) reports to have an accuracy of 0.84.

These models are not interesting only for their promising accuracy, but for what they are trained to measure. The authors of the original paper introducing the Financial PhraseBank, Malo et al. (2014) state about annotating the training data: "the annotators were asked to consider the sentences from the view point of an investor only; i.e. whether the news may have positive, negative or neutral influence on the stock price. As a result, sentences which have a sentiment that is not relevant from an economic or financial perspective are considered neutral". This means that the datasets annotations, and therefore any language model trained with this dataset, should assess a sentence positive only if the information might have a positive effect on the stock price. In other words, the FinBERT models should only be expected to assess the sentiment of value-relevant information, disregarding statements that are too vague or non-material to be relevant for firm valuation.

3 Motivation, Research Questions & Hypotheses

3.1 Motivation

Understanding qualitative management disclosures and the information value that different types of soft information have for investors and financial markets in general has not been an easy task for academics to tackle. The management disclosures of a firm are inherently intriguing, offering valuable insights and crucial context beyond its financial performance. A compelling story provided by a company can wield a transformative influence on its valuation, transcending the numerical realm of financials to shape perceptions, build investor confidence, and ultimately steer the trajectory of market value.

In more specific terms, understanding management disclosures and their impact on stock returns can provide us with important insights on several aspects. Firstly, there are obvious policy implications. Regulators try to encourage companies to provide truthful and relevant information in their disclosures. Understanding the value-relevance of different disclosure types, as well as the discretionary strategic decisions that managers take when communicating with market participants, is crucial for shaping the regulatory environment to allow for open communication, but safeguard investors from biased or flat-out deceptive communication. Secondly, there are implications to how managers should communicate to investors. Managers seeking to enhance or uphold their firm's information environment naturally find value in understanding what makes management disclosures effective and informative. Thirdly, given the influence of management incentives on disclosures and the presence of premeditated statements conveying biased or imperfect information, investors are eager to discern the genuinely informative aspects within management disclosures and understand their significance. Answering these questions has not been easy. From an academic perspective, one problem has been suboptimal data availability. In addition to the mentioned hindrance of lack of annotated corpora, for example earnings calls have not have a similar history of availability than most other types of firm disclosures. Regulation Fair Disclosure -rule allowed public access to US firms' earnings conference calls only since 2000 (SEC, 2000). Another issue has been academics relying too much on rudimentary methodology.

Even though many of newer methods have been proven to be reliable and useful in other scientific fields, including many areas that similarly to finance also have specialized language (such as medicine), these advancements have not been generally adopted by academics in the accounting and finance disciplines. In this thesis, I will attempt to decouple the connection between value-relevant soft information and the linguistic tone with which this information is being communicated. To the best of my knowledge,

isolating and assessing both components simultaneously has not been attempted before with any financial soft information.

In terms of information processing aspect, it is obviously interesting to learn how quickly and effectively financial markets process information and how it is incorporated into share prices. As Martineau (2021) finds, the PEAD has all but disappeared for the "hard" earnings surprise information. Price et al. (2012) studied the relationship between soft information in earnings calls and PEAD from 2004 to 2007. Since then, however, markets have had a much longer time to learn about the information content in earnings calls, and the market reactions to this information may be very different than what they used to be.

Finally, while there have been some interesting developments in utilizing LLMs in financial content analysis effectively, the real-life performance of these new models is still mostly yet to be tested. Different FinBERT models have shown some very promising performance when used for classifying sentences in the Financial PhraseBank dataset. Whether this performance translates to ability in measuring textual sentiment in the real world, and more specifically earnings calls, is still very much unknown.

3.2 Research questions and hypotheses

I will first study the potential benefits of using the LLM-based FinancialBERT to extract soft information in quarterly earnings calls. As mentioned, the previously used dictionary methods all suffer from lack of sophistication, as this approach is unable to understand context even on a basic level. On the other hand, FinancialBERT has shown very promising performance in classifying the Financial PhraseBank dataset (Hazourli 2022b). I therefore hypothesize that FinancialBERT can be used to extract sentiment from earnings calls, and that it outperforms Henry's dictionary-method, which was found to beat the alternative dictionary from McDonald & Loughran (2011) by Price et al. (2012). Consequently, my first research question and hypotheses are as follows:

Research Question 1: Can LLM-based sentiment analysis tools be used to measure the soft information in firms' quarterly earnings calls in general, and can they provide improved performance over the previous methods?

H1a: The call sentiment as measured with FinancialBERT strongly predicts the cumulative abnormal returns during the earnings announcement window (from day -1 to day 1 around the call).

H1b: The FinancialBERT-based approach outperforms the previous method of using Henry's dictionary-based approach, both in statistical and economic significance.

As mentioned, the human-made annotation of Financial PhraseBank, the dataset with which FinancialBERT has been trained on, is based on classifying whether a sentence would be likely to have either positive, negative, or no impact on the firm's stock price (Malo et al. 2014). In contrast to this, Henry's wordlist was created on the idea that it is not what is being said, but rather how it is said (Henry 2008). The choice of words included in this dictionary therefore reflects this decision. It's reasonable to assume that both the actual message and how it is being conveyed could provide relevant information to market participants. While abnormal tone in firm disclosures has been determined to be bad news (e.g. X. Huang et al. 2014, Blau et al. 2015, and Baginski et al. 2018), there is still strong evidence that managers' do provide soft information that is relevant for firm value in their disclosures (e.g. Price et al. 2012; Hennig et al. 2023). There have not been robust ways to decouple these two phenomena from each other, and academics studying either phenomenon might often have captured a combination of stock price reactions related to these distinct factors. This leads me to my next research question and hypotheses: As FinancialBERT has been designed to extract value-relevant information from financial texts, that is now something that can be controlled for in the statistical test. Thus, Henry's wordlist should then act not as a substitute for the sentiment measure, but rather as a complement, capturing only the "how" of information disclosure, or the pure linguistic tone. When controlling for value-relevant information with FinancialBERT sentiment, I expect the impact that this tone measure captures to be analogous to the "abnormal tone" that some academics have studied and found to impact share prices negatively. Another question is the timing of this negative impact of tone. Previous literature has shown that it would come with a delay (X. Huang et al. 2014; Baginski et al. 2018), although apart from Hennig (2023) not much is known about it in earnings calls context. I theorize that these previous findings do not reflect the full truth about abnormal tone; I find it likely that in the existing literature the initial positive impact is capturing more of the reaction to the value-relevant information, thus concealing the negative impact. I therefore hypothesize that the negative impact of linguistic tone would behave largely in the same way as the reactions to value-relevant information in its timing; The impact is at least partly incorporated into share prices immediately, but fully only after the extended reaction period. I also expect this phenomenon to be most prominent in the calls' presentation section, as this is the pre-prepared portion of the call over which the managers have the biggest control over (Price et al. 2012; Blau et al. 2015).

Research Question 2: Can Henry's wordlist approach be used as a purer measure of linguistic tone, when controlling for value-relevant soft information with FinancialBERT?

H2a: Henry's dictionary -measure captures different phenomenon than FinancialBERT in the calls' presentation section when both are included in a regression.

H2b: Controlling for value-relevant information with FinancialBERT sentiment, Henry's tone measure has a negative impact on cumulative abnormal returns, which is partly occurring during the initial 3-day reaction period, and fully after the extended reaction period (from trading days 2 to 60 from call date).

Finally, if the attempt to isolate the effects of value-relevant information and linguistic tone is successful, there remains the question of how quickly markets incorporate the value-relevant information into share prices. Here there are a few possible alternatives: The stock prices tend to react to soft information at least partly immediately, but there is also a substantial part of the reaction that has historically happened during an extended period (e.g. Price et al. 2012; X. Huang et al. 2014; Baginski et al. 2018; Hennig et al. 2023). However, what we know specifically about the market reactions to overall level of soft information, as opposed to only the abnormal part, is that this type of information has taken some time to be reflected in the stock prices (Price et al. 2012). Thus, my third research question and hypothesis:

Research Question 3: Does the earnings call -based drift on soft information during the post-earnings announcement period persist between 2008 and 2021?

H3: The sentiment of the call, for both presentation and the Q&A sections, is associated positively with the abnormal returns and is a significant predictor for abnormal returns during the extended stock price reaction period.

3.3 Contribution

In this thesis I aim to provide three main contributions to the existing literature. Firstly, I am presenting the methodological contribution of utilizing LLM-based approach in financial sentiment analysis. These recently developed models show promising results when assessing their test data from the training dataset, but they have yet to be shown to work properly with real-world data and beyond measuring the sentiment of individual sentences. If found to function as intended, this new approach could completely change the status quo on how financial textual content analysis research is conducted. In a discipline where almost all the work is done with relatively unsophisticated methodology, a shift to a more modern approach could also help alleviate doubts of the quality of academic research in the field. In short, demonstrating the performance of LLM-based models in real-world

application can help the academic field to produce more robust and higher quality research in the future, which would help the field of financial textual analysis to become more recognized and credible area in financial academic literature.

Secondly, I use this methodology to study the stock price reactions to explicitly value-relevant information in earnings calls, i.e. information that one would expect to have a stock price reaction. To the best of my knowledge, there is a dearth of research on this precise phenomenon. The existing literature is almost exclusively focused on studying the impact of linguistic tone either explicitly or implicitly, rather than the value-relevant soft information content in earnings calls, or any other medium. This value-relevant textual information is the actual closest soft information counterpart to the hard information of earnings figures in earnings announcements. Thus, one could argue that this phenomenon of "what" is being said during earnings calls is an even more profound question than "how" it is said, to advance our understanding of the soft information released around firm earnings announcements.

Finally, I introduce an important distinction between the two phenomena of sentiment – value-relevant information in financial text – and linguistic tone. What follows from allowing this separation of the two effects is not only a better understanding of the less researched impact of value-relevant information, but also a better understanding of the impact of the linguistic tone. As far as I am aware, only Blau et al. (2015) study linguistic tone in earnings calls with a method that controls for value-relevant information in any way, but they do not show a direct connection between that "abnormal" tone and stock returns. Hence, a part of this study's contribution is also to be the first paper to study the impact of linguistic tone on stock returns in a way that is not as prone to capturing the actual information content of the statements.

4 Data and methods

4.1 Sample description

I study the effects that soft information in quarterly earnings calls has on stock returns between the years 2008 and 2021. For the sample selection process, I follow the methodology of Price et al. (2012), with some adjustments. I filter the companies in the three major US stock exchanges (NASDAQ, American Stock Exchange, and NYSE) with the following criteria: I exclude REITs, ADRs, closed end funds and units by filtering out all other CRSP share codes than 10, 11 and 12. I also exclude other financials (SIC 6000-6999) and utilities (SIC 4900-4949).

I collect a pseudorandom sample of firm-quarter observations by sorting the remaining companies into quartiles by both size, measured by market capitalization, and Book-to-Market (B/M) equity, to achieve a representative sample of observations from firms with different characteristics. The size quartile sorting is based on market capitalization breakpoints for NYSE firms, made available by Kenneth French online (French 2023). The observations are sorted into size quartiles based on their 1-quarter lagged market capitalizations. The B/M equity quartile sorting is done with breakpoints calculated from all firms of the three sample stock exchanges in CRSP Compustat merged database with similar share code limitations. I calculate the B/M equities for each observation with a twice-lagged book equity and once-lagged market equity values. Before collecting the pseudorandom sample, I exclude observations where the firm's share price is below USD 1, or the market capitalization is below USD 5 million. I also leave out observations with a negative book equity value. Finally, I drop observations without sufficient data for this study, such as missing controls data items, analyst estimates data, or insufficient stock return data for either determining the abnormal returns model (pre-announcement returns) or studying the returns following the announcement. I then randomly pick 5 firms from each of the 16 size-B/M portfolios for each of the 56 quarters. I collect the conference call transcripts for this sample from two sources: Refinitiv Eikon and S&P Capital IQ. After some final data exclusions based on data availability and eliminating a small number of outliers, I am left with a total of 4,287 individual firm-quarter observations, which is my final sample.

4.2 Methodology for sentiment analysis

In this study, I measure textual soft information with two different approaches: First, I recreate the tone measure used by Price et al. (2012), which is based on Henry's (2008) wordlist. Second, I create a new sentiment measure using the machine learning model FinancialBERT developed by Hazourli

(2022b). While I first test how these two measures succeed in measuring the overall soft information content in a call, I continue with an important distinction between two types of soft information in earnings calls: tone and sentiment. I define tone as the linguistic tone of language, or how things are said, which I go on to measure with Henry's wordlist approach. I define sentiment as the value-relevant information content in the statements themselves, or what is being said, which I measure with FinancialBERT.

For Henry's tone measure, the approach is relatively simple. I use the two wordlists developed by Henry (2008), one of positive words and one of negative words, which are presented in Appendix 1. For any given text, I count the total number of matching words found for both lists, to get total numbers of positive words and negative words in the text. Price et al. (2012) calculate two different tone ratios from these numbers. The first ratio is the number of positive words divided by the number of negative words. The second ratio is the number of positive words minus the number of negative words, divided by the total number of either positive or negative words:

$$TONE = \frac{(Positive\ words - Negative\ words)}{(Positive\ words + Negative\ words)} \quad (1)$$

While Price et al. (2012) test both ratios and state that their results are almost identical with either measure, their main study focuses on the second measure. This is also the ratio that I use in this study.

The second approach utilizing Hazourli's language model differs greatly from this word count method. For this method, I first split the text into individual sentences using Natural Language Toolkit", a commonly used NLP-library. I then feed each sentence into Hazourli's model (Hazourli 2022a), obtaining classification for each sentence into one of the three following categories: "positive", "negative", or "neutral". I present some examples of classified sentences along with Henry's tone measure information in Appendix 2. I calculate two different ratios for total sentiment (SENT) of a text. First, the ratio of the difference between the number of positive and negative sentences, scaled by total number of sentences (positive, negative, or neutral):

$$SENT_1 = \frac{(Positive\ sentences - Negative\ sentences)}{Total\ sentences} \quad (2)$$

Secondly, not unlike with the Henry's tone measure, I calculate the ratio of the difference between positive and negative sentences, divided by the sum of positive and negative sentences:

$$SENT_2 = \frac{(Positive\ sentences - Negative\ sentences)}{(Positive\ sentences + Negative\ sentences)} \quad (3)$$

While clearly somewhat different measures, there exists a theoretical justification for using either one of the measures, that relates to what type of sentences the ones classified as "neutral" are. As a reminder to the reader, the FinancialBERT model is trained to do its classification task with a dataset where sentences were annotated as either positive, negative, or neutral depending on whether the sentence "may have positive, negative or neutral influence on the stock price". The sentences marked as neutral might have been classified as such by Hazourli's model for either one of the following two reasons: First, if the news is relevant to investors, but is not expected to cause positive or negative stock price reactions. For example, sentences might state how the company's operations are on track with the current expectations. Such a piece of information would definitely be relevant for the firm's valuation, but without giving it any reason to change. Second possible reason for a sentence to be classified as neutral is if it has no value-relevant information to begin with. The calls often include exclamations such as "Good morning!", or statements about the call structure itself, such as the managers asking the operators to move to the next question. In the case of the first explanation, it would make more sense to scale the difference between positive and negative sentences by the total number of sentences, allowing the "neutral" information to also influence the sentiment score. With the second case, however, it makes more sense to ignore the neutral statements altogether and scale the difference only with the sentences guaranteed to have information value, as is done in the second formula. As one might expect, in the collected transcripts neither explanation is completely accurate; There exist a fair amount of both types of neutral sentences. However, upon inspecting the tendencies of neutrally tagged statements, they seem to be value-irrelevant more frequently. I therefore use the formula 3 sentiment measure for the remainder of this thesis.

4.3 Empirical approach and regressions

To measure how the sentiment or tone information in earnings calls affects stock returns, I study the impact of this information on (cumulative) abnormal stock returns around and after the earnings calls. I collect the earnings call dates and times from the transcript files. I adjust for any calls taking place after the stock market has already closed by changing the call date to be the next trading day. I define abnormal return AR as the return for stock j on day t as the excess return for that stock relative to the expected return measured with Fama-French 3-factor + momentum model. I extract the abnormal returns from WRDS Event Study Tool. This tool estimates the following formula for the model returns:

$$R_{FFM3+MOM} = R_f + \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \beta_4 MOM \quad (4)$$

Where $R_{FFM3+MOM}$ represents the model returns, R_f is the risk-free rate, R_m is the market return, SMB is the returns of small over large stocks, HML is the returns of high B/M over low B/M stocks, and MOM represents the returns of past winners over past losers. For this estimation, I use the following parameters: The estimation window is set at 100 trading days, with a 60-day gap to the event date (day 0). The minimum number of valid observations for the estimation period is 70, which all quarter-firm observations fulfil without exceptions. This means that the above formula is estimated for the firms using daily return data from dates between -170 and -70. The tool then calculates the abnormal returns in the following manner:

$$AR_{j,t} = R_{j,t} - R_{FFM3+MOM,j,t} \quad (5)$$

Where $AR_{j,t}$ is the abnormal returns for firm j in day t, $R_{j,t}$ is the returns for firm j on day t, and $R_{FFM3+MOM,j,t}$ is the expected (model) returns for firm j in day t. These daily abnormal returns can then be added together from the desired date range to reach the cumulative abnormal return for that period. Similarly to Price et al. (2012) I assess the cumulative abnormal returns mainly within two time windows: Firstly in the immediate call window, days -1 to 1 (where day 0 is the date of the earnings call), and secondly during the longer-term reaction period from days 2 to 60 of the earnings call period. I call the first period the "initial reaction period", and the second one the "extended reaction period". The (-1, 1) period is meant to capture the immediate stock price reaction (as well as measure the overall effectiveness of the tested tone or sentiment tool), whereas the (2, 60) period measures the PEAD period, to study how the soft information is incorporated into share prices over a longer time period. As mentioned, while the hard information PEAD has all but disappeared (Martineau 2021), soft information could still take longer time to be fully reflected in the share prices. Price et al. (2012) find the (2, 60) measurement period useful in their study of tone information in earnings calls, showing an existing drift for soft information in that period. Therefore, the formulas for the cumulative abnormal returns are as follows:

$$CAR(-1,1)_j = \sum_{t=-1}^{1} AR_{j,t} \quad (6)$$

$$CAR(2,60)_j = \sum_{t=2}^{60} AR_{j,t} \quad (7)$$

I present preliminary analysis of the cumulative abnormal returns by sorting the observations into quintiles first based on the Henry's tone measure (following Price et al. 2012), and then based on the sentiment measure by Hazourli. I conduct the difference of means test and the Wilcoxon rank test for medians between the top and bottom quintiles. I perform this analysis for both tone measures, and for both time periods.

Diverging from Price et al.'s (2012) methodology, I use the analyst surprise instead of the simple "random walk" earnings surprise to control for the hard information in the earnings announcement. Whereas the random walk earnings surprise is calculated based on the difference between the earnings per share (EPS) in the current quarter and the same quarter in the previous year, the analyst surprise is calculated from the difference between the actual quarterly EPS and the consensus analyst estimate. This is a method that is recommended by for example Martineau (2021), who finds that the random walk surprise is a much noisier measure of the earnings surprise. While this decision limits my sample only to firms with some analyst following, I deem the improved surprise measure a beneficial benefit from the trade-off, as the hard earnings information should now be captured as accurately as possible by the surprise variable, instead of being inadvertently captured in any way by my tone/sentiment measure. I retrieve the actual EPS figures, as originally reported during the earnings announcement, as well as the latest consensus EPS estimates before the announcement from I/B/E/S database. I calculate the analyst surprise with the following formula:

$$SURP_j = \frac{(EPS_j - E(EPS_j)}{Price_{j,t-5}} * 100 \qquad (8)$$

Where j is the firm and a is the earnings announcement date. In other words, I calculate the difference between the announced EPS and the expected EPS (analyst consensus estimate just before the earnings announcement) scaled by the firm's stock price 5 trading days before the announcement.

Using the CARs and the earnings surprises, I then conduct a visual inspection of the cumulative abnormal returns for the period between 10 days before the earnings call to 60 days after the call, using the same tone/sentiment quintiles as in the tests for differences of means and medians. I study the CARs for the top and bottom quintiles of the tone/sentiment portfolios, for both Henry's tone measure (following Price et al. 2012) and FinancialBERT sentiment. I also visually examine the performance of portfolios formed by both the tone and sentiment measure terciles, as well as earnings surprise terciles.

Next, again following Price et al.'s (2012) methodology, I perform a regression analysis for the full sample, for both the initial and extended reaction periods' cumulative abnormal returns, firstly using the Henry's tone measurement:

$$CAR_j = \gamma_{0,i} + \gamma_{1,i} SURP_{i,j} + \gamma_{2,i} TONE_{i,j} + CONTROLS_j + \varepsilon_j \quad (9)$$

where CAR is the cumulative abnormal return for an earnings call j, for the defined time period. TONE is the Henry's tone measure as defined above and SURP is the analyst surprise as defined above. Similarly to Price et al. (2012), CONTROLS include total word counts of the earnings call transcript's presentation and Q&A parts (Pres. word count and Q&A word count) to control for the amount of information in the call, SIZE is the log of the firm's market capitalization at the end of the previous quarter, BM is the book-to-market equity at the end of the previous quarter, PROFITABILITY is the profitability measured as net income divided by total assets multiplied by 100, LEVERAGE is calculated as total liabilities divided by total assets multiplied by 100, VOLUME is the log of total share trading volume on call day, VOLATILITY is the standard deviation of the firm's daily returns in the 90-day time period of (-100, -10) from the earnings call date. ANALYSTS measures the analyst coverage for the firm and is calculated as the log of the number of analysts covering the firm most recently before the call. DECLARATION is a dummy variable that is set as 1, if the firm declares dividend in the 3-day window around the call between dates (-1, 1), the variable is set to 0 if there is no dividend declaration during this period. The data for the control variables is acquired from CRSP and Compustat, with the exception of the data for analyst following and surprises, which is retrieved from I/B/E/S. I control for heteroscedasticity as well as fixed firm effect with clustering by firm. Finally, for some of the additional analysis in Chapter 6, I divide the sample based on the share of institutional ownership, the data for which I retrieve from Thomson/Refinitiv database.

I then conduct the same regression analysis, but this time with the Hazourli-based sentiment measure:

$$CAR_j = \gamma_{0,i} + \gamma_{1,i} SURP_{i,j} + \gamma_{2,i} SENT_{i,j} + CONTROLS_j + \varepsilon_j \quad (10)$$

Finally, I conduct a regression with both Henry's tone measure as well as Hazourli's sentiment measure, to differentiate between the value relevant information (Hazourli's sentiment) and the linguistic tone (Henry's tone):

$$CAR_j = \gamma_{0,i} + \gamma_{1,i} SURP_{i,j} + \gamma_{2,i} TONE_{i,j} + \gamma_{3,i} SENT_{i,j} + CONTROLS_j + \varepsilon_j \quad (11)$$

The earnings conference calls generally consist of two parts: Management presentation and a subsequent Q&A part. During the first part, management of the company usually presents the quarterly earnings results and provides

additional commentary that has been prepared in advance. During the Q&A part of the call, analysts have the opportunity to ask, and have the management answer, questions about the company. These two parts are therefore very distinct from each other. The management has a better chance to control the tone and information during the presentation part, whereas the Q&A sessions puts them on the spot to respond to questions about what the analysts deem relevant to ask about the firm. Price et al. (2012) note that the presentation part very often just reiterates the information contained in the earnings press release, and they therefore conduct their analysis separating the two parts from each other, to prevent their measure from just capturing the tone of the earnings press release. In this study, I make the same distinction, and separate the Presentation and Q&A parts from each other, calculating the sentiment and tone measures separately for the two parts.

I expect to see a stronger statistical and economic impact in cumulative abnormal returns for the FinancialBERT sentiment measure in the initial 3-day call window than for the Henry's tone measure, for both the presentation and Q&A parts of the call. For the extended period, I similarly expect to see the FinancialBERT sentiment predict abnormal returns more strongly than Henry's tone. However, as there has been numerous studies documenting the negative impact of inflated or residual tone to stock returns, I expect the Henry's tone measure to negatively predict the cumulative abnormal returns when included in the regression with FinancialBERT sentiment within both time windows. There is also reason to believe that this abnormal tone impact would be more prominent in the presentation part of the call, as that where managers should be more able to control the linguistic tone.

5 Main analysis results

5.1 Descriptive statistics and correlations

Table 1. Descriptive statistics and correlations

This table shows the descriptive statistics and correlations for the cumulative abnormal returns, earnings surprise, and tone and sentiment measure figures for the total sample. CAR(-1, 1) is the cumulative abnormal return within the three-day period around the earnings call (where earnings call takes place on day 0) and CAR(2,60) is the cumulative abnormal return over the subsequent drift period, from days 2 to 60 after the earnings call. Both cumulative abnormal returns are presented as a percentage (scaled by a hundred) and are the abnormal returns relative to the expected return, which is estimated using the Fama-French 3-factors + momentum -model. SURP is the analyst surprise. The SURP measure is calculated as the difference between the consensus quarterly EPS estimate and the actual announced quarterly EPS, divided by the share price 5 days before the earnings announcement, times 100. TONE is the Henry's tone measure for the call transcript, a ratio of positive and negative words, calculated as (positive - negative) / (positive + negative). SENT is the sentiment measure based on Hazourli's FinancialBERT's sentence classifications, calculated as (positive - negative) / (positive + negative).

	CAR -1, 1	CAR 2, 60	SURP	TONE	SENT
Panel A: Descriptive statistics					
Mean	-0.10	-0.07	0.01	0.60	0.66
Min	-55.77	-268.70	-23.31	-0.25	-0.27
P25	-4.55	-11.16	-0.10	0.51	0.55
P50/Median	0.12	-0.31	0.06	0.63	0.69
P75	4.52	11.16	0.27	0.72	0.79
Max	57.75	227.53	12.90	0.95	1.00
Std.Dev	8.99	23.47	1.60	0.17	0.19
N	4287	4287	4287	4287	4287
Panel B: Correlations					
CAR -1, 1	1.00				
CAR 2, 60	0.03	1.00			
SURP	0.17	-0.03	1.00		
TONE	0.14	-0.08	0.15	1.00	
SENT	0.20	-0.08	0.17	0.82	1.00

Panel A of Table 1 shows the descriptive statistics for the cumulative abnormal returns, earnings surprise, Henry's tone measure, and FinancialBERT sentiment measure for the full sample of 4287 observations. The cumulative abnormal returns for the initial reaction period appear relatively symmetric around 0. The mean cumulative abnormal return during this period is -0.10%, while the median for the same period is 0.12%. The cumulative abnormal returns range between the minimum and maximum values of -

55.77% and 57.75% respectively. However, the 25th and 75th percentile figures of -4.55% and 4.52%, as well as a standard deviation of 8.99, show an effect that is much more muted, yet still sizeable enough to show the impact of all the new information presented to the public during the quarterly earnings announcements.

The figures for the extended reaction period (trading days 2 to 60 relative to the earnings call) similarly show cumulative abnormal returns that tend to be close to zero; The mean and median for this period's CARs are -0.07% and -0.31% respectively. On this much longer time period, we see quite expectedly also larger deviations in the cumulative abnormal returns. The cumulative abnormal returns of my sample have the 25th and 75th percentiles at -11.16% and 11.16% respectively, and a standard deviation of 23.47. These figures show again a relatively symmetric distribution of abnormal returns around zero, even though the numbers do show a somewhat longer left tail with a minimum value of -268.70 in contrast with the maximum value of 227.53.

The mean of the analyst earnings surprise (SURP) is very close to zero, at 0.01, with a median of 0.06, suggesting a sample that is fairly equally distributed between positive and negative earnings surprises. Half of the earnings surprises in the sample are between -0.10 and 0.27, which are the 25th and 75th percentiles of the earnings surprise. The minimum value for the earnings surprise is -23.31, and the maximum value is 12.90. The standard deviation for the earnings surprise is 1.60.

Moving onto the tone and sentiment measures, it is good to remember that as results of their formulas, both measures are restricted to values between -1 and 1. A tone or sentiment score of a zero would indicate a "neutral" tone or sentiment, with the same amount of positive and negative words (Henry's tone measure) or sentences (FinancialBERT sentiment measure). For Henry's tone measure (TONE), the sample has a mean (median) of 0.60 (0.63), indicating an overall positive tone across the earnings calls. The minimum and maximum values in the sample are -0.25 and 0.95, showing a range between calls with a somewhat negative tone, and calls with extremely positive tone. The 25th percentile is at 0.51, and the 75th percentile at 0.72, indicating that most calls are dispersed within a relatively small range of optimism around the median.

The sentiment scores (SENT) derived from FinancialBERT's classifications paint a qualitatively similar picture: the mean (median) score is 0.66 (0.69), indicating again that a typical call sentiment is clearly optimistic. The full range of scores for the sample is from -0.27 and 1, qualitatively similar, this time reaching the highest possible score with no negatively classified sentences in a single earnings call. The 25th and 75th percentiles for the sentiment measures are 0.55 and 0.79 respectively, again similarly in line with Henry's tone measure, if not just slightly more positive. The call tones being positive on average is most likely a result of managers' incentives to portray

their firms in a more positive light. This is consistent with for example Kartik et al.'s (2007) model of strategic communication in which the used language is inflated in equilibrium, as well as with Davis & Tama-Sweet's (2012) findings that managers do indeed use more optimistic language in presence of higher incentives to do so.

While the descriptive statistics do not show clear differences between the tone and sentiment measures, the correlation analysis in panel B of table 1 reveals some differences between the two. When looking at the first column, correlations between the announcement window CARs and other variables, I find a very small correlation between the abnormal returns of the initial and extended reaction periods, suggesting that there would likely not be a strong connection between the returns on these two time periods. I do however find some indication on how the earnings surprise, tone, and sentiment might be linked to the announcement period returns: The earnings surprise has a correlation of 0.17 with the initial period cumulative abnormal returns, indicating – quite expectedly – a positive relationship the two variables. Here I also find the first hints about the differences between the Henry's tone measure and FinancialBERT sentiment: TONE has a correlation coefficient with the CAR(-1, 1) of 0.14, suggesting a connection that is weaker than the one the earnings surprise has with the returns. On the other hand, FinancialBERT sentiment measure has a correlation of 0.20 with the initial reaction period abnormal returns, suggesting a stronger connection with the immediate returns than either TONE or SURP.

Analysing the correlations for the extended reaction period returns, I find slightly negative correlations of -0.08 for both Henry's tone and FinancialBERT sentiment, and an almost zero correlation of -0.03 correlation for the earnings surprise. These correlations do not suggest a strong post-earnings announcement drift, based on neither the hard nor soft information. If anything, these correlations would seem to hint at the opposite: a potential reversal of the initial returns.

Finally, the table shows us that the Henry's tone and FinancialBERT sentiment are highly correlated with each other, with a correlation coefficient of 0.82. Judging from this number alone, it is not possible to say if one or these measures would outperform the other. It is hardly surprising for these two measures to have a strong and positive correlation coefficient, but this correlation is still not perfect.

5.2 Portfolio-level analysis of soft information

Table 2 shows the means and medians for different tone quintiles as measured with the Henry's tone measure (TONE), as well as the tests for the differences in means and medians between the top and bottom quintiles. The portfolio sorting has been done first using the full earnings call, then only the managers' presentation section, and finally using only the Questions and Answers -section of the call. In general, the results for these different call sections are relatively similar. The top tone quintile has a significantly greater mean and median CAR than the bottom quintile during the initial reaction period. These results are statistically significant for the total call, as well as for both distinct call sections at 1% level.

Table 2. Differences of means and medians - TONE

This table shows the means and medians for quintiles sorted with the Henry's tone measure (TONE), as well as the test of differences of means and medians for the top and bottom portfolios. Of the quintiles, portfolio 1 is the one with the lowest tone measure, and 5 the one with the highest tone measure. There are three tone figures in the table: "Total TONE" for the tone of a full call, "Presentation TONE" for the tone in the managers' presentation section of a call, and "Q&A TONE" for the tone in the Questions and Answers -section of a call. I show means and medians for the cumulative abnormal returns (CAR) of two distinct time periods: "CAR(-1, 1)" for the initial 3-day call period around the earnings call (call date = 0), and "CAR(2, 60)" period for the drift period.

1 = low, 5 = high		Total TONE quintiles		Pres. TONE quintiles		Q&A TONE quintiles	
		CAR(-1, 1)	CAR(2, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 1)	CAR(2, 60)
1	Mean	-1.92 %	3.08 %	-1.81 %	3.76 %	-1.65 %	0.87 %
	Median	-1.55 %	2.58 %	-1.33 %	2.36 %	-1.11 %	0.78 %
2	Mean	-0.98 %	1.56 %	-0.89 %	1.06 %	-0.85 %	0.94 %
	Median	-0.50 %	-0.24 %	-0.50 %	1.03 %	-0.15 %	0.44 %
3	Mean	-0.07 %	-0.76 %	0.16 %	-0.44 %	0.02 %	0.30 %
	Median	0.05 %	-0.76 %	0.14 %	-1.00 %	0.01 %	0.23 %
4	Mean	0.80 %	-1.43 %	0.51 %	-1.42 %	0.86 %	-1.12 %
	Median	0.79 %	-0.87 %	0.79 %	-1.01 %	0.58 %	-1.27 %
5	Mean	1.70 %	-2.79 %	1.54 %	-3.30 %	1.13 %	-1.33 %
	Median	1.51 %	-1.97 %	1.49 %	-2.38 %	1.05 %	-1.46 %
Mean Q5-Q1		3.62 %	-5.87 %	3.35 %	-7.06 %	2.78 %	-2.20 %
t-statistic		-8.09	5.05	-7.56	6.11	-6.21	1.88
p-value		0.00	0.00	0.00	0.00	0.00	0.06
Wilcoxon rank-sum test							
Median Q5-Q1		3.06 %	-4.55 %	2.83 %	-4.73 %	2.16 %	-2.24 %
z-statistic		-8.52	5.69	-8.18	6.19	-6.29	2.49
p-value		0.00	0.00	0.00	0.00	0.00	0.01

For this initial earnings call window, both the mean and median CARs increase consistently from the lowest tone quintiles to the highest. This holds true for the tone of the total call, presentation, as well as the Q&A section. For the total call tone, the mean (median) CAR in the bottom quintile is -1.92% (-1.55%), compared to the CAR in the top quintile of 1.70% (1.51%), resulting in a difference of 3.62% (3.06). The results for the two separate call section tones are qualitatively similar. The portfolios based on presentation section's tone have a mean (median) CAR of -1.81% (-1.33%) in the bottom quintile and 1.54% (1.49%) in the top quintile, a difference of 3.35% (2.83%). The Q&A portfolios have only slightly smaller differences with a mean (median) CAR of -1.65% (-1.11%) in the bottom quintile and 1.13% (1.05%) in the top quintile, resulting in a difference of 2.78% (2.16%). All the differences listed above are statistically significant at 1% level. These results are in line with the findings by Price et al. (2012), showing the impact of soft information in earnings call in the initial earnings call time frame: A higher call tone does seem to result in a higher stock return during the initial reaction period.

However, if we shift our focus to the returns in the extended reaction period (the trading days from 2 to 60 relative to the call), the picture seems very different. For the total call tone and the two separate call section tones, we find the top tone quintile to have a significantly smaller returns than the bottom quintile. For the total call quintiles, it is the bottom portfolio that has a strongly positive mean (median) CAR of 3.08% (2.58%), compared to the top portfolio's -2.79% (-1.97%). This results in the mean (median) difference in CAR of -5.87% (-4.55%) between the top and bottom quintiles. In other words, in this extended period, more positive tone results in a clearly smaller cumulative abnormal return. For the presentation call section, the results are similar, if not even stronger: The mean (median) return for the bottom portfolio is 3.76% (2.36%), compared to the top portfolio's -3.30% (-2.38%), resulting in a difference of -7.06% (-4.73%). All these results for the total call and presentation section quintiles are significant at 1% level. For the Q&A section, the results point to an impact to the similar direction, albeit smaller and with a slightly weaker statistical significance. For the Q&A tone quintiles, the bottom portfolio has a mean (median) CAR of 0.87% (0.78%), compared to the top portfolio's -1.33% (-1.46%). This leads to the top portfolio having a mean CAR that is 2.20% smaller than the bottom portfolio's (significant at 10% level), and a median that is 2.24% smaller (significant at 5% level).

Overall, the results from this portfolio approach point at two different takeaways: Firstly, the results do show the positive impact of a higher tone in the initial 3-day earnings call time frame, consistent with previous studies; A higher tone does indeed seem to be associated with a higher CAR, for the total call as well as both the presentation and Q&A section. Secondly, during the extended stock reaction period the results suggest an opposite effect, a

potential reversal, where a higher tone predicts smaller stock returns than a lower one.

Table 3 shows the corresponding quintile means and medians, for the initial and extended stock reaction period, but this time sorted using the FinancialBERT sentiment (SENT). In qualitative terms, the results using FinancialBERT do not differ from the results obtained using the Henry's tone measure: Higher tone quintile is associated with a higher CAR for the initial reaction period, but a smaller CAR for the extended reaction period. However, there are some smaller differences in the figures.

Table 3. Differences of means and medians - SENT

This table shows the means and medians for quintiles sorted with the FinancialBERT sentiment measure (SENT), as well as the test of differences of means and medians for the top and bottom portfolios. Of the quintiles, portfolio 1 is the one with the lowest sentiment measure, and 5 the one with the highest sentiment measure. There are three sentiment figures in the table: "Total SENT" for the sentiment of a full call, "Presentation SENT" for the sentiment in the managers' presentation section of a call, and "Q&A SENT" for the sentiment in the Questions and Answers -section of a call. I show means and medians for the cumulative abnormal returns (CAR) of two distinct time periods: "CAR(-1, 1)" for the initial 3-day call period around the earnings call (call date = 0), and "CAR(2, 60)" period for the drift period.

1 = low, 5 = high		Total TONE quintiles		Pres. TONE quintiles		Q&A TONE quintiles	
		CAR(-1, 1)	CAR(2, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 1)	CAR(2, 60)
1	Mean	-2.75 %	2.91 %	-2.48 %	3.45 %	-2.30 %	0.73 %
	Median	-1.95 %	1.53 %	-1.77 %	2.71 %	-1.77 %	0.12 %
2	Mean	-0.79 %	1.12 %	-0.90 %	0.39 %	-0.97 %	1.79 %
	Median	-0.48 %	1.02 %	-0.57 %	0.23 %	-0.47 %	0.87 %
3	Mean	-0.14 %	-0.32 %	0.00 %	0.31 %	0.16 %	-0.62 %
	Median	0.10 %	-0.32 %	0.27 %	-0.46 %	0.58 %	-0.18 %
4	Mean	1.05 %	-0.87 %	0.81 %	-1.29 %	1.02 %	-1.17 %
	Median	0.76 %	-1.08 %	0.76 %	-1.36 %	0.72 %	-0.94 %
5	Mean	2.15 %	-3.18 %	2.09 %	-3.20 %	1.60 %	-1.07 %
	Median	1.76 %	-2.44 %	1.55 %	-2.32 %	1.27 %	-1.63 %
Mean Q5-Q1		4.89 %	-6.08 %	4.57 %	-6.65 %	3.90 %	-1.79 %
t-statistic		-10.95	5.02	-10.25	5.59	-8.67	1.48
p-value		0.00	0.00	0.00	0.00	0.00	0.14
Wilcoxon rank-sum test							
Median Q5-Q1		3.70 %	-3.97 %	3.33 %	-5.03 %	3.05 %	-1.75 %
z-statistic		-11.20	5.60	-10.06	5.93	-8.88	2.58
p-value		0.00	0.00	0.00	0.00	0.00	0.01

For the initial reaction period CARs, the differences in means and medians calculated using FinancialBERT are across the board higher, hinting that the measure might be better at capturing the soft information in the earnings calls. For the total call, the mean (median) CAR for the lowest

sentiment quintile is -2.75% (-1.95%), compared to the highest quintile's 2.15% (1.76%), leading to a difference of 4.89% (3.70%). These differences are larger than the ones measured with Henry's tone measure (3.62% difference of means and 3.06% difference of medians). The same applies for the separate sections' sentiment: For the presentation section, the difference of means and medians for the top and bottom quintiles are 4.57% and 3.33% respectively, compared to 3.35% and 2.83% obtained with sorting by Henry's measure. For the Q&A section the difference of means and medians are 3.90% and 3.05% respectively, compared to 2.78% and 2.16% from the Henry's tone measure sorts.

For the extended reaction period, the results are more mixed, although the big picture remains unchanged: Higher sentiment seems to predict lower CARs during the extended reaction period. The difference of means between the top and bottom quintiles for the total call sentiment is -6.08%, slightly larger than with Henry's measure, but the difference of medians is -3.97%, which is somewhat less extreme than with Henry's measure. For the presentation section, the differences compare to Henry's measure the other way around: Difference of means is less extreme, but still notably large -6.65%, and the difference of medians is just slightly bigger than with Henry's tone quintiles at -5.03%. For the Q&A section of the call, economic significance is smaller for both measures when using FinancialBERT, with a difference of means (medians) at -1.79% (-1.75%). Again, the results suggest that there would not be a momentum-like drift for the extended reaction period, but the opposite: a reversal of the initial reaction.

Next, following the methodology of Price et al. (2012) I move to visually inspect the abnormal cumulative returns for the portfolios. I conduct this study for the top and bottom quintiles first formed with the Henry's tone measure, and then with FinancialBERT measure. After this, I look at portfolios based on either of these two measures combined with their earnings surprises.

Looking at Figure 1, we can see very clearly what the tests for differences of means and medians hinted at: While there is a strong initial stock return reaction to the direction of the tone measure (higher tone resulting in higher returns and vice versa), there seems to be both swift and continuous reversal, where the observations with the lowest tones clearly outperform the ones with the highest tones. Quite interestingly, this reversal effect looks to be strong and long-lasting enough to completely shift cumulative abnormal returns in favour of the firms with low call tones. While the initial reaction causes the CARs for the top (bottom) tone quintile to peak at 1.5% (-1.9%) on day 2, the CARs converge at around the 30-day mark and continue this reversal even past their starting point from before the earnings announcement. After the full 60-trading days after the earnings call, the CAR for the top (bottom) tone quintile stands at -1.4% (1.3%). Indeed, based on this analysis, it is the low tone observations that yield the highest CARs compared to the low

tone observations. This might hint that the strategic communication explanation is appropriate for the earnings calls; The Henry's tone measure might be capturing the inflated tone of the call, which has been shown to be a negative sign for the firm's stock returns.

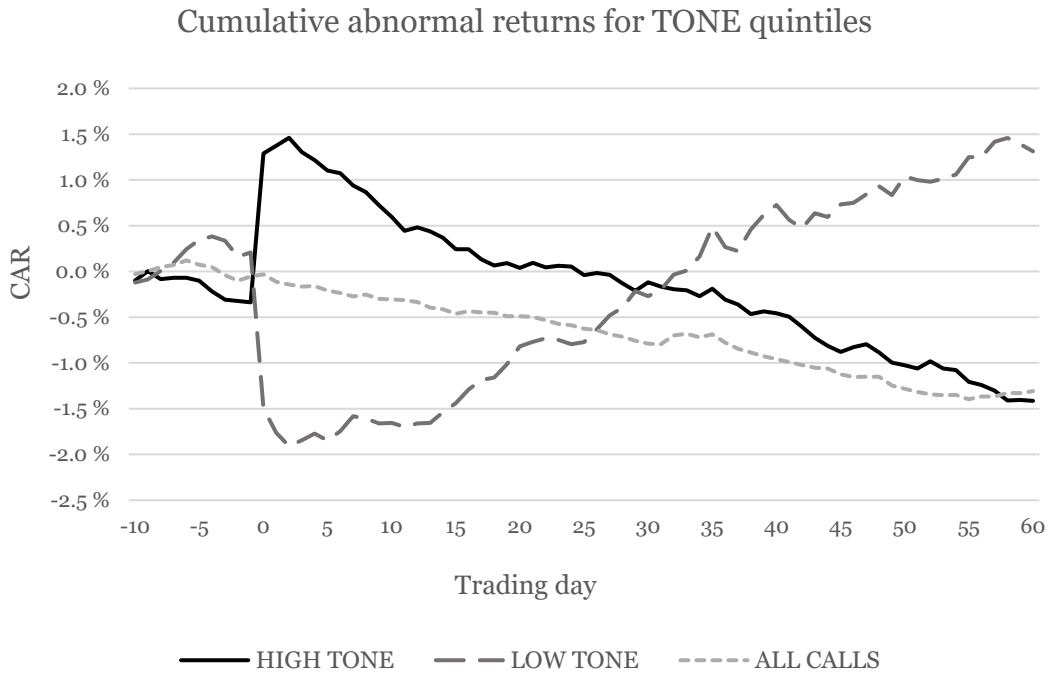

Figure 1. Cumulative abnormal returns on the top and bottom TONE quintiles. This figure depicts the cumulative abnormal returns for the top and bottom quintiles sorted by the total call's TONE. TONE is the tone of the call as measured with Henry's (2008) tone measure. The average CARs of the total sample is also shown. The abnormal returns are depicted on the y-axis. The x-axis shows the trading days relative to the date of the earnings call from 10 days before the call (-10) to 60 days after it (60).

Figure 2 shows a similar analysis as figure 1, but this time on quintiles based on FinancialBERT sentiment. Qualitatively the results seem relatively similar. However, when sorted to quintiles with FinancialBERT, the initial stock return reactions are stronger, and the reversal effect both slower and somewhat weaker in the full study period: The CAR for the top (bottom) sentiment quintile peaks on day 2 at 1.7% (-2.4%). These returns then start to converge and reach each other on day 43. On trading day 60, the cumulative abnormal return for the top (bottom) sentiment quintile is -1.4% (0.6%). While the reversal effect is slightly smaller than when measured with Henry's measure, these results still leave some questions unanswered. To understand the impact of call tone on the stock returns better, I next move to assess these phenomena together with the observations' earnings surprises.

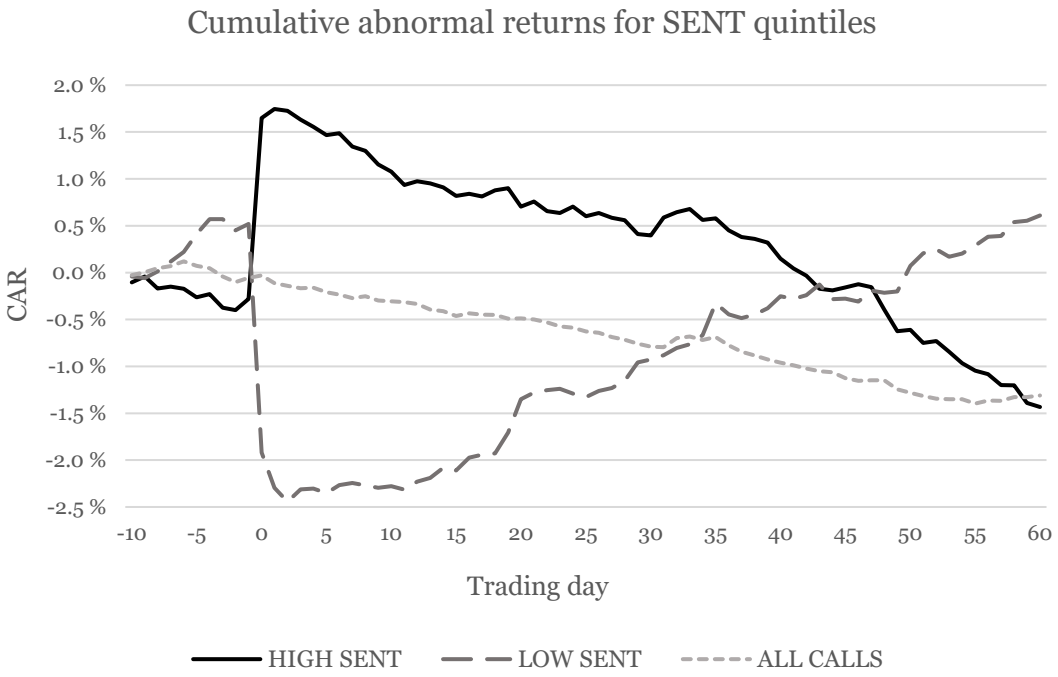

Figure 2. This figure depicts the cumulative abnormal returns for the top and bottom quintiles sorted by the total call's SENT. SENT is the sentiment of the call as measured with FinancialBERT-based measure. The average CARs of the total sample is also shown. The abnormal returns are depicted on the y-axis. The x-axis shows the trading days relative to the date of the earnings call from 10 days before the call (-10) to 60 days after it (60).

Figure 3 shows the cumulative abnormal returns of four different portfolios, based on the earnings surprise and the Henry's tone measure: "low surprise, low tone" -portfolio includes observations that rank in the bottom terciles both in earnings surprise as well as earnings call tone. Conversely, "high surprise, high tone" portfolio includes observations in the top tercile for both earnings surprise and tone. "Low surprise, high tone" -portfolio includes observations that are in the top earnings surprise tercile but in the bottom tone tercile, and "high surprise, low tone" -portfolio includes observations that are in the top earnings surprise tercile but bottom call tone tercile.

Based on this approach, the sign of the initial stock price reaction seems to be determined by the earnings surprise, The initial reaction is quite unsurprisingly the highest for the high surprise, high tone -portfolio, and the lowest for the low surprise, low tone -portfolio, with CARs on day 1 of 4.0% and -3.7% respectively. The high surprise, low tone portfolio has a positive initial reaction, with a day 1 CAR of 1.5%. Its counterpart, low surprise, high tone portfolio has a negative reaction, with a corresponding CAR at -2.8%. It would appear that while the initial reaction is largely defined by the earnings surprise, the earnings call tone can either amplify or mitigate the size of this reaction.

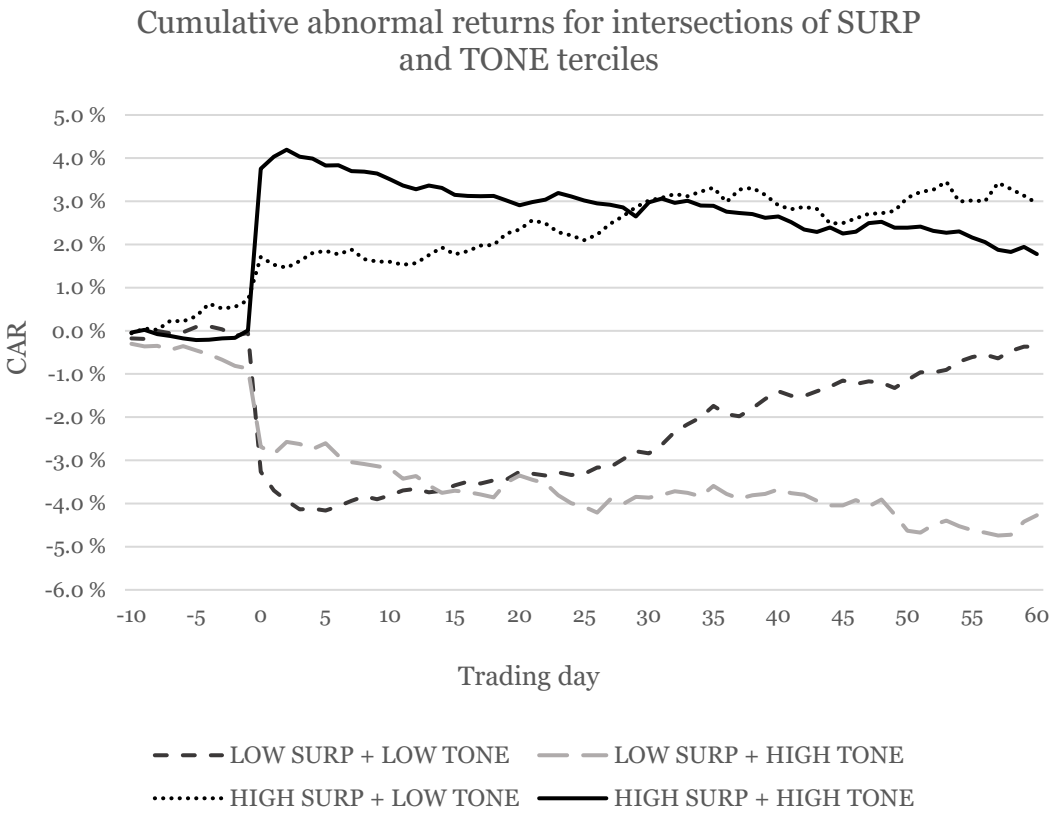

Figure 3. This figure depicts the cumulative abnormal returns for four portfolios, formed based on the earnings surprise as well as the Henry's tone measure. The "LOW SURP" ("HIGH SURP") observations are the earnings announcements with earnings surprise in the bottom (top) tercile. The observations in the "LOW TONE" ("HIGH TONE") are the earnings calls with tone measure in the bottom (top) tercile of the sample. These two measures are used to form the four portfolios, for all the different combinations of these terciles. TONE is the tone of the call as measured with Henry's (2008) tone measure. The average CARs of the total sample is also shown. The abnormal returns are depicted on the y-axis. The x-axis shows the trading days relative to the date of the earnings call from 10 days before the call (-10) to 60 days after it (60).

During the extended reaction period, another phenomenon can be seen from the figure. For the portfolios where the tone amplified the initial stock return reaction, there is a clear reversal effect, that seems to persist for the full duration of the study period, for all the 60 trading days after the earnings call. The cumulative abnormal returns for the high surprise, high tone portfolio reverts from its peak of 4.2% on day 2 to 1.8% on day 60. For the low surprise, low tone portfolio, this reversal happens from ca. -4.2% on day 5 to -0.4% on day 60.

For the two portfolios with contradictory earnings surprises and tones, a different effect is found: The initial dampened reaction to the earnings surprise is followed by a drift to the same direction. For the high surprise, low tone portfolio, this drift brings the CAR from 1.5% on day 1 to 3.0% at the end

of the extended period. For the low surprise, hight tone portfolio, the CAR drifts from -2.8% to -4.3% in the same period.

Figure 4 shows a similar analysis but using the FinancialBERT sentiment measure instead of the Henry's tone measure. Again, the results are qualitatively speaking very similar, with the same initial stock price reaction dominated by the earnings surprise effect, and either amplified or mitigated by the sentiment measure, leading then to either a reversal or a drift, based on whether the sentiment of the earnings call exaggerated or dampened the initial reaction. And in the same way, the observations with a higher call tone perform worse relative to their lower tone counterpart.

The similarity of these two measures with the portfolio approach is unsurprising, as the Henry's tone and FinancialBERT sentiment measures have a relatively high correlation, and figures 3 and 4 are based on sorting the observations into terciles.

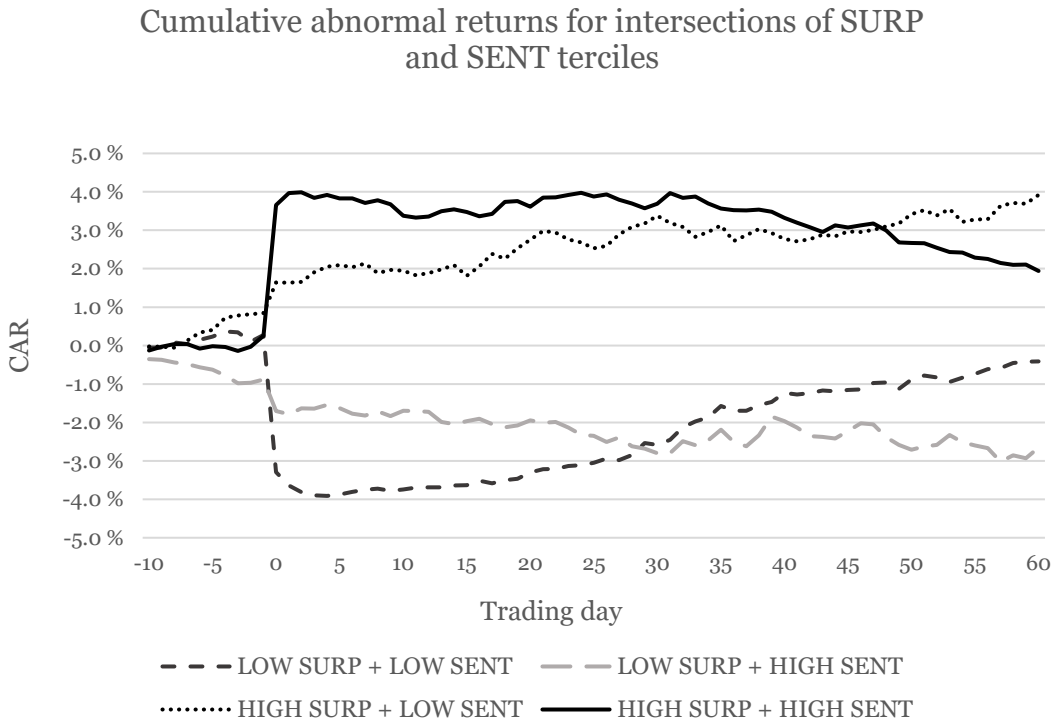

Figure 4. This figure depicts the cumulative abnormal returns for four portfolios, formed based on the earnings surprise as well as the FinancialBERT sentiment measure. The "LOW SURP" ("HIGH SURP") observations are the earnings announcements with earnings surprise in the bottom (top) tercile. The observations in the "LOW SENT" ("HIGH SENT") are the earnings calls with the FinancialBERT sentiment measure in the bottom (top) tercile of the sample. These two measures are used to form the four portfolios, for all the different combinations of these terciles. SENT is the sentiment of the call as measured with FinancialBERT. The average CARs of the total sample is also shown. The abnormal returns are depicted on the y-axis. The x-axis shows the trading days relative to the date of the earnings call from 10 days before the call (-10) to 60 days after it (60).

5.3 Comparison of sentiment measures

To answer the first research question on the relative performance between FinancialBERT sentiment and Henry's tone, I study how well these measures capture the economic impact in abnormal stock returns during the initial earnings call reaction period. Table 4 shows the regression summary results for the dependent variable CAR(-1, 1), the cumulative abnormal returns for the initial reaction period. Models in Panel A are different model configurations with the Henry's tone measure. Models in Panel B are corresponding model configurations, only with FinancialBERT sentiment measure instead of the Henry's tone measure.

What unites all these models is the highly significant – on a 1% statistical significance level – coefficients on both the earnings surprise as well as the Henry's tone and FinancialBERT sentiment measures. What can be seen from the two models with full controls (on the right-most column of Panels A and B), is that the tones or sentiments of both the management presentation section as well as the Q&A section are very highly statistically significant. What can also be seen from the results is that the FinancialBERT-based sentiment measure seems outperform Henry's tone measure in capturing the soft information content in the earnings calls. The statistical significance of the SENT measures clearly beats the corresponding TONE measure's significance for the whole call, management presentation section, as well as the Q&A section in every single model configuration. Comparing the results for the two configurations with all control variables, the model with the FinancialBERT sentiment measure has an adjusted R-squared of 0.067, clearly outperforming the corresponding model with TONE measure, which has an adjusted R-squared of 0.050.

The economic significance is also clearly stronger for the FinancialBERT sentiment -measures. Comparing the two models with full controls, an increase of one standard deviation in the presentation section's FinancialBERT sentiment results in a 1.00% higher abnormal cumulative return in the initial reaction period. For the presentation section's TONE, one standard deviation -increase results in 0.73% higher returns. The captured stock price return impact is therefore 38% larger for the FinancialBERT sentiment than for the TONE. For the Q&A section, the corresponding impact is 1.02% for the FinancialBERT sentiment and 0.70% for TONE, which means that for a single standard deviation change, FinancialBERT measure is capturing a return impact that is a full 47% larger than the TONE measure.

These results show strong support for my hypotheses 1a and 1b, showing that a sentiment measure that is calculated from an LLM-based FinancialBERT-model handily outperforms the previous methodology of using the simpler dictionary-based Henry's tone.

Table 4. Regression results on the cumulative abnormal returns during the initial reaction period on earnings surprise and earnings call tone/sentiment

This table shows the regression statistics on the cumulative abnormal returns for the initial stock reaction period CAR(-1, 1). Panel A shows regressions on different configurations of earnings surprise and Henry's tone measure (TONE). Panel B shows regression results on similar configurations, but for the FinancialBERT sentiment (SENT) instead of TONE. SURP is the analyst surprise, as defined in the data and methodology section of this paper. TONE total, TONE pres. and TONE Q&A are the Henry's tone measures for the full earnings call, the managers' presentation section, and the Q&A section, respectively. SENT total, SENT Pres., and SENT Q&A are the corresponding FinancialBERT sentiment measures for the same textual items. "Pres. word count" and "Q&A word count" are the number of words in the presentation and Q&A sections of the call, respectively, in thousands. SIZE is the log of firm market capitalizations, as at the end of previous quarter. B/M is the book-to-market equity at the end of the previous quarter. PROFITABILITY is the profitability of the firm, calculated as the net income divided by total assets times 100. LEVERAGE is total liabilities divided by total assets times 100. VOLUME is the log of total share trading volume on the day of the earnings call. VOLATILITY is the standard deviation of daily returns, for the 90 trading days on the time period ending 10 days before the earnings call. ANALYSTS is the log of number of analysts covering the firm immediately before the earnings call. DECLARATION is 1 if the firm has declared dividends in the 3-day window around the earnings call, and zero otherwise. The dependent variable CAR(-1, 1) is in percentages. T-statistics for the coefficient estimates are in parentheses.

*** = p < 0.01, ** = p < 0.05, * = p < 0.1

Panel A: TONE models	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**
SURP	0.860***	0.883***	0.899***	0.867***	0.836***
	(6.19)	*(6.30)*	*(6.54)*	*(6.23)*	*(6.25)*
TONE total	6.381***				
	(7.20)				
TONE pres.		4.036***		2.658***	3.340***
		(5.84)		*(3.34)*	*(4.22)*
TONE Q&A			5.355***	3.643***	4.190***
			(6.14)	*(3.62)*	*(4.11)*
Pres. word count					-0.202*
					(-1.72)
Q&A word count					0.021
					(0.22)
SIZE					0.201
					(1.28)
B/M					0.955**
					(2.53)
PROFITABILITY					0.103**
					(2.44)
LEVERAGE					0.004
					(0.68)
VOLUME					-0.556***
					(-3.52)
VOLATILITY					0.371***
					(2.71)
ANALYSTS					0.537*
					(1.72)
DECLARATION					0.303
					(0.90)
Constant	-3.940***	-2.516***	-3.353***	-3.900***	-0.610
	(-7.00)	*(-5.66)*	*(-5.97)*	*(-6.64)*	*(-0.39)*
Observations	4287	4287	4287	4287	4287
R-squared	0.043	0.039	0.039	0.042	0.053
Adjusted R-squared	0.042	0.038	0.039	0.041	0.050

Table 4 (continued)

Panel B: SENT models	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**
SURP	0.797***	0.837***	0.884***	0.823***	0.774***
	(5.83)	*(6.06)*	*(6.38)*	*(5.94)*	*(5.85)*
SENT total	8.408***				
	(10.22)				
SENT pres.		5.561***		3.758***	4.349***
		(8.00)		*(4.96)*	*(5.76)*
SENT Q&A			6.329***	4.430***	4.979***
			(8.67)	*(5.52)*	*(6.17)*
Pres. word count					-0.234**
					(-2.03)
Q&A word count					0.082
					(0.88)
SIZE					0.199
					(1.29)
B/M					1.012***
					(2.65)
PROFITABILITY					0.133***
					(3.16)
LEVERAGE					0.005
					(0.77)
VOLUME					-0.582***
					(-3.75)
VOLATILITY					0.348**
					(2.59)
ANALYSTS					0.439
					(1.41)
DECLARATION					0.316
					(0.95)
Constant	-5.630***	-3.725***	-4.315***	-5.496***	-1.789
	(-9.95)	*(-7.73)*	*(-8.43)*	*(-9.60)*	*(-1.17)*
Observations	4287	4287	4287	4287	4287
R-squared	0.059	0.049	0.050	0.057	0.070
Adjusted R-squared	0.058	0.049	0.050	0.057	0.067

5.4 Regression results for the extended reaction period

I then move my attention to the extended stock return period. To study whether the post-earnings announcement drift for the earnings calls' soft information – as described by Price et al. (2012) – persists in this new sample period, I regress the cumulative abnormal returns from day 2 to 60 relative to the earnings call date on the call tone/sentiment measures. Table 5 shows the regression results for similar model configurations than Table 4, only this time with the extended period stock returns as the dependent variable.

Looking at these results, and specifically the two models with full controls for TONE and FinancialBERT sentiment (on the right-most column of Panels A and B), there is clear evidence of a return reversal on the presentation section tone or sentiment. As was suggested by the portfolio-level study with tests of differences of means and medians, as well as with the visual inspection of the portfolio returns, the call tone or sentiment negatively predicts the stock returns between day 2 and 60 from the call date. Looking at the results, it is also evident that this predictive value is focused specifically on the management presentation section of the call. In the models with full controls, I find the TONE and SENT measures of the Q&A section to be insignificant, but the same tone and sentiment measures to be highly significant (at 1-% level) for the presentation section.

The economic impact of this phenomenon is also sizeable: For the presentation section's TONE, a single standard deviation increase in the tone predicts 2.19% lower abnormal returns in the extended return period. A single standard deviation increase in the FinancialBERT sentiment of the presentation section predicts 2.30% lower returns for the same period. These results, together with the similar evidence from the portfolio-level study, are at odds with my hypothesis on the persistence of the post-earnings announcement drift on soft information in earnings calls. To the contrary, there is clear indication on a reversal on the tone or sentiment of the presentation section of the call.

Table 5. Regression results on the cumulative abnormal returns during the extended reaction period on earnings surprise and earnings call tone/sentiment

This table shows the regression statistics on the cumulative abnormal returns for the extended stock reaction period CAR(2, 60). Panel A shows regressions on different configurations of earnings surprise and Henry's tone measure (TONE). Panel B shows regression results on similar configurations, but for the FinancialBERT sentiment (SENT) instead of TONE. SURP is the analyst surprise, as defined in the data and methodology section of this paper. TONE total, TONE pres. and TONE Q&A are the Henry's tone measures for the full earnings call, the managers' presentation section, and the Q&A section, respectively. SENT total, SENT Pres., and SENT Q&A are the corresponding FinancialBERT sentiment measures for the same textual items. "Pres. word count" and "Q&A word count" are the number of words in the presentation and Q&A sections of the call, respectively, in thousands. SIZE is the log of firm market capitalizations, as at the end of previous quarter. B/M is the book-to-market equity at the end of the previous quarter. PROFITABILITY is the profitability of the firm, calculated as the net income divided by total assets times 100. LEVERAGE is total liabilities divided by total assets times 100. VOLUME is the log of total share trading volume on the day of the earnings call. VOLATILITY is the standard deviation of daily returns, for the 90 trading days on the time period ending 10 days before the earnings call. ANALYSTS is the log of number of analysts covering the firm immediately before the earnings call. DECLARATION is 1 if the firm has declared dividends in the 3-day window around the earnings call, and zero otherwise. The dependent variable CAR(2, 60) is in percentages. T-statistics for the coefficient estimates are in parentheses.

*** = p < 0.01, ** = p < 0.05, * = p < 0.1

Panel A: TONE models	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**
SURP	-0.251	-0.224	-0.368	-0.233	-0.030
	(-0.55)	*(-0.49)*	*(-0.81)*	*(-0.51)*	*(-0.06)*
TONE total	-11.063***				
	(-4.62)				
TONE pres.		-10.402***		-11.198***	-10.070***
		(-5.88)		*(-5.32)*	*(-4.85)*
TONE Q&A			-5.110**	2.103	1.995
			(-2.11)	*(0.73)*	*(0.67)*
Pres. word count					-0.077
					(-0.26)
Q&A word count					-0.009
					(-0.04)
SIZE					-1.247***
					(-3.31)
B/M					2.432*
					(1.72)
PROFITABILITY					-0.252*
					(-1.79)
LEVERAGE					-0.017
					(-0.98)
VOLUME					0.562
					(1.46)
VOLATILITY					-0.633
					(-1.36)
ANALYSTS					2.490***
					(3.02)
DECLARATION					0.399
					(0.44)
Constant	6.578***	6.138***	3.032*	5.338***	2.669
	(4.36)	*(5.45)*	*(1.96)*	*(3.40)*	*(0.57)*
Observations	4287	4287	4287	4287	4287
R-squared	0.007	0.010	0.002	0.010	0.022
Adjusted R-squared	0.006	0.009	0.002	0.009	0.019

Table 5 (continued)

Panel B: SENT models	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**
SURP	-0.231	-0.207	-0.384	-0.212	0.025
	(-0.50)	*(-0.45)*	*(-0.85)*	*(-0.46)*	*(0.05)*
SENT total	-10.037***				
	(-4.61)				
SENT pres.		-9.850***		-10.557***	-9.954***
		(-5.98)		*(-5.64)*	*(-5.42)*
SENT Q&A			-3.597*	1.738	1.275
			(-1.89)	*(0.81)*	*(0.58)*
Pres. word count					-0.025
					(-0.08)
Q&A word count					-0.061
					(-0.23)
SIZE					-1.276***
					(-3.37)
B/M					2.585*
					(1.86)
PROFITABILITY					-0.270*
					(-1.92)
LEVERAGE					-0.017
					(-0.97)
VOLUME					0.582
					(1.52)
VOLATILITY					-0.585
					(-1.27)
ANALYSTS					2.604***
					(3.17)
DECLARATION					0.567
					(0.63)
Constant	6.528***	6.341***	2.328*	5.646***	3.031
	(4.52)	*(5.73)*	*(1.83)*	*(4.12)*	*(0.68)*
Observations	4287	4287	4287	4287	4287
R-squared	0.007	0.010	0.002	0.010	0.023
Adjusted R-squared	0.006	0.010	0.001	0.010	0.020

5.5 Regression with both tone and sentiment measures

Before moving to a more in-depth interpretation of the above regression results, I also run regressions for both study periods with both Henry's tone and FinancialBERT sentiment measures together, results of which can be seen in table 6. As a remainder, relatively few studies in the financial sentiment analysis field have attempted to clearly explain their definition of "tone" or "sentiment", which could in many cases be attributed to the existing range of methodologies implicitly determining the definition. However, as has already been noted, Henry (2008) originally designed her wordlist to measure the "how" things are being said, instead of "what" exactly is said. On the other hand, the Financial PhraseBank dataset has been annotated based on whether the information in a sentence is likely to cause the stock price to react either positively or negatively, setting a higher bar for the actual information value in a text than Henry's tone measure. Instead of treating tone or sentiment as a "general positivity/negativity" of a text, as much of the existing literature does, I test whether the Henry's wordlist and FinancialBERT can be used complementarily: Henry's wordlist measuring the tone, or "how" things are said, and FinancialBERT measuring the more objective value-relevance of the text. I test this idea, defined in research question 2 and hypotheses 2a and 2b, by running regression on the initial and extended stock reaction period, but this time including both Henry's tone measure and FinancialBERT sentiment in the same models.

What is revealed is that for the initial stock reaction period, Henry's tone measure has a negative impact on stock returns, when I control for value-relevant information with the SENT variables. A one standard deviation change in the presentation section's TONE is associated with a -0.57% change in cumulative abnormal returns for the initial period. This impact is significant at a 5-% level. The TONE for the Q&A section is not significantly different from zero. The measure of value-relevant information, FinancialBERT sentiment (SENT) is associated with a clear positive impact. For the presentation and Q&A sections, an increase of one standard deviation is associated with a 1.41% and 0.98% increase in the cumulative abnormal returns respectively, at a 1-% significance level. While the economic impact for the Q&A section's sentiment is practically the same with or without including the TONE measures, the sentiment of the presentation section's impact on cumulative abnormal returns is 0.40 percentage points, or a full 40% higher than without the inclusion of Henry's tone measures. Overall, these results support my hypothesis 2a, that the two measures – despite being relatively strongly correlated – do indeed capture different phenomena. The abnormal or residual tone of management disclosures has been shown to predict negative returns, and the TONE measure seems to be capturing this impact at least to some extent.

Table 6. Regression results on the cumulative abnormal returns during the initial and extended reaction period on earnings surprise and earnings call tone and sentiment

This table shows regression results on the cumulative abnormal returns for both initial and extended stock reaction periods. SURP is the analyst surprise, as defined in the data and methodology section of this paper. TONE total, TONE pres. and TONE Q&A are the Henry's tone measures for the full earnings call, the managers' presentation section, and the Q&A section, respectively. SENT total, SENT pres., and SENT Q&A are the corresponding FinancialBERT measures for the same textual items. "Pres. word count" and "Q&A word count" are the numbers of words in the presentation and Q&A sections of calls in thousands. SIZE is the log of firm market capitalizations, as at the end of previous quarter. B/M is the book-to-market equity at the end of the previous quarter. PROFITABILITY is the profitability of the firm, calculated as the net income divided by total assets times 100. LEVERAGE is total liabilities divided by total assets times 100. VOLUME is the log of total share trading volume on the day of the earnings call. VOLATILITY is the standard deviation of daily returns, for the 90 trading days on the time period ending 10 days before the earnings call. ANALYSTS is the log of number of analysts covering the firm immediately before the earnings call. DECLARATION is 1 if the firm has declared dividends in the 3-day window around the earnings call, and zero otherwise. The dependent variables CARs are in percentages. T-statistics for the coefficient estimates are in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1

Panel A: CAR(-1, 1)	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**	**CAR(-1, 1)**
SURP	0.803***	0.842***	0.879***	0.827***	0.772***
	(5.90)	*(6.12)*	*(6.35)*	*(6.02)*	*(5.85)*
TONE total	-3.584**				
	(-2.43)				
TONE pres.		-1.831		-2.839**	-2.637**
		(-1.60)		*(-2.40)*	*(-2.26)*
TONE Q&A			0.883	0.682	0.973
			(0.78)	*(0.57)*	*(0.81)*
SENT total	11.022***				
	(7.99)				
SENT pres.		6.937***		5.713***	6.083***
		(5.95)		*(5.00)*	*(5.40)*
SENT Q&A			5.895***	4.382***	4.785***
			(6.30)	*(4.56)*	*(5.00)*
Pres. word count					-0.243**
					(-2.11)
Q&A word count					0.091
					(0.98)
SIZE					0.203
					(1.33)
B/M					0.932**
					(2.43)
PROFITABILITY					0.136***
					(3.25)
LEVERAGE					0.005
					(0.74)
VOLUME					-0.581***
					(-3.75)
VOLATILITY					0.327**
					(2.41)
ANALYSTS					0.412
					(1.33)
DECLARATION					0.270
					(0.81)
Constant	-5.195***	-3.529***	-4.562***	-5.457***	-1.697
	(-8.86)	*(-7.20)*	*(-7.63)*	*(-8.66)*	*(-1.08)*
Observations	4287	4287	4287	4287	4287
R-squared	0.060	0.050	0.050	0.059	0.071
Adjusted R-squared	0.059	0.049	0.050	0.058	0.068

Table 6 (continued)

Panel B: CAR(2, 60)	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**	**CAR(2, 60)**
SURP	-0.221	-0.191	-0.362	-0.200	0.021
	(-0.48)	*(-0.42)*	*(-0.80)*	*(-0.44)*	*(0.04)*
TONE total	-5.841				
	(-1.38)				
TONE pres.		-5.690*		-6.366**	-4.043
		(-1.93)		*(-2.04)*	*(-1.32)*
TONE Q&A			-3.797	0.961	1.439
			(-1.24)	*(0.30)*	*(0.44)*
SENT total	-5.776				
	(-1.50)				
SENT pres.		-5.573**		-6.108**	-7.293***
		(-2.01)		*(-2.20)*	*(-2.66)*
SENT Q&A			-1.730	1.879	0.999
			(-0.72)	*(0.76)*	*(0.41)*
Pres. word count					-0.038
					(-0.13)
Q&A word count					-0.047
					(-0.18)
SIZE					-1.269***
					(-3.36)
B/M					2.462*
					(1.74)
PROFITABILITY					-0.264*
					(-1.88)
LEVERAGE					-0.017
					(-0.99)
VOLUME					0.584
					(1.52)
VOLATILITY					-0.618
					(-1.33)
ANALYSTS					2.564***
					(3.13)
DECLARATION					0.497
					(0.55)
Constant	7.236***	6.952***	3.387**	5.872***	3.192
	(4.69)	*(5.93)*	*(2.13)*	*(3.64)*	*(0.68)*
Observations	4287	4287	4287	4287	4287
R-squared	0.008	0.011	0.002	0.011	0.023
Adjusted R-squared	0.007	0.010	0.002	0.010	0.020

It is when turning the attention to the extended reaction period, however, that the results appear particularly interesting. Firstly, the TONE of the presentation section and the sentiment of the Q&A section are not statistically significantly different from zero during the extended reaction period, indicating that all the information contained in these measures are, contrary to hypothesis 3, incorporated into share prices with no delay. Secondly, while these tone or sentiment measures are not statistically significant in this period, the FinancialBERT sentiment of the presentation section is strongly and negatively significant at 1-% level. A single standard deviation change in this sentiment measure predicts a 1.68% change in abnormal returns to the opposite direction. This finding is clearly contrary to hypothesis 3 predicting a

drift in abnormal returns. Not only does this result fail to support that hypothesis, but it suggests a completely opposite effect: There is a meaningful reversal effect on the abnormal returns brought from the earnings call presentation section's sentiment. Not only that, but the coefficient estimates for the two different reaction periods suggest that this reversal is roughly equally large to the initial reaction impact, potentially sweeping away all the entirety of abnormal returns observed during the initial period.

To study this phenomenon more closely, and to confirm the total impact of different types of information in earnings calls, I run additional regressions, this time for the complete, combined initial and extended stock reaction period, results of which can be seen in Table 7. While the results do get somewhat noisier with this time window – trying to capture the abnormal return impact on the full reaction period in these models results in consistently smaller R-squared scores than when regressing only on either initial or extended reaction period – the same phenomena that were found in the earlier analysis are evident. What can be seen from the model in the right-most column is that in the long term there indeed does seem to be only two statistically significant types of information in the earnings calls: Firstly, the Henry's tone of the presentation section has a negative impact (as was seen in Table 6), and the FinancialBERT sentiment of the Q&A section has a positive impact on the abnormal returns. Both effects are significant at a 5-% level. The sentiment of the presentation section is not statistically significantly different from zero.

5.6 Discussion about the results

Having now conducted the main analyses for this paper, it is now possible to thoroughly assess the research questions and hypotheses based on the results. First topic of discussion is research question 1, about whether LLM-based sentiment analysis tools ,such as FinancialBERT, can be used to measure the soft information in firms' earnings calls. I find strong support for both hypotheses 1a, that call sentiment strongly predicts the cumulative abnormal returns for firms during the initial earnings reaction period, as well as H1b, that FinancialBERT-based measure outperforms Henry's tone measure. Using the FinancialBERT sentiment measure instead of Henry's tone improves the regression model's explanatory power from an R-squared of 0.053 to 0.070, a roughly 30% improvement. For the Q&A section of the call, the results show that the FinancialBERT sentiment is positively associated with the full reaction period returns at a 1-% significance level, showing a clearly stronger statistical significance than Henry's tone's 10-% level in the same period. It seems clear that despite the relatively good performance of the Henry's tone, the LLM-based sentiment measure clearly outperforms it.

Table 7. Regression results on the cumulative abnormal returns during the full reaction period on earnings surprise and earnings call tone and sentiment

This table shows the regression statistics on the cumulative abnormal returns for the full stock price reaction period, from day -1 to day 60 relative to the earnings call date. The different model configurations show the regression results for different combinations of the TONE and SENT measures. SURP is the analyst surprise, as defined in the data and methodology section of this paper. TONE pres. and TONE Q&A are the Henry's tone measures for the managers' presentation section, and the Q&A section, respectively. SENT pres. and SENT Q&A are the corresponding FinancialBERT sentiment measures for the same textual items. "Pres. word count" and "Q&A word count" are the number of words in the presentation and Q&A sections of the call, respectively, in thousands. SIZE is the log of firm market capitalizations, as at the end of previous quarter. B/M is the book-to-market equity at the end of the previous quarter. PROFITABILITY is the profitability of the firm, calculated as the net income divided by total assets times 100. LEVERAGE is total liabilities divided by total assets times 100. VOLUME is the log of total share trading volume on the day of the earnings call. VOLATILITY is the standard deviation of daily returns, for the 90 trading days on the time period ending 10 days before the earnings call. ANALYSTS is the log of number of analysts covering the firm immediately before the earnings call. DECLARATION is 1 if the firm has declared dividends in the 3-day window around the earnings call, and zero otherwise. The dependent variable CARs are in percentages. T-statistics for the coefficient estimates are in parentheses.

*** = p < 0.01, ** = p < 0.05, * = p < 0.1

	CAR(-1, 60)	**CAR(-1, 60)**	**CAR(-1, 60)**	**CAR(-1, 60)**
SURP	0.806	0.799	0.627	0.792
	(1.61)	*(1.59)*	*(1.28)*	*(1.58)*
TONE pres.	-6.730***		-9.204***	-6.680**
	(-3.00)		*(-2.71)*	*(-2.00)*
TONE Q&A	6.186*		1.643	2.412
	(1.96)		*(0.47)*	*(0.69)*
SENT pres.		-5.605***	-0.396	-1.211
		(-2.84)	*(-0.13)*	*(-0.40)*
SENT Q&A		6.253***	6.261**	5.784**
		(2.60)	*(2.29)*	*(2.15)*
Pres. word count	-0.279	-0.259		-0.281
	(-0.88)	*(-0.82)*		*(-0.89)*
Q&A word count	0.012	0.021		0.045
	(0.04)	*(0.08)*		*(0.16)*
SIZE	-1.046**	-1.077***		-1.066**
	(-2.54)	*(-2.60)*		*(-2.58)*
B/M	3.388**	3.596***		3.394**
	(2.44)	*(2.64)*		*(2.45)*
PROFITABILITY	-0.149	-0.137		-0.128
	(-0.99)	*(-0.91)*		*(-0.85)*
LEVERAGE	-0.013	-0.012		-0.013
	(-0.70)	*(-0.66)*		*(-0.69)*
VOLUME	0.006	0.000		0.002
	(0.01)	*(0.00)*		*(0.01)*
VOLATILITY	-0.262	-0.237		-0.291
	(-0.54)	*(-0.49)*		*(-0.60)*
ANALYSTS	3.028***	3.043***		2.976***
	(3.35)	*(3.38)*		*(3.31)*
DECLARATION	0.701	0.883		0.767
	(0.74)	*(0.93)*		*(0.81)*
Constant	2.060	1.242	0.415	1.495
	(0.41)	*(0.26)*	*(0.24)*	*(0.30)*
Observations	4287	4287	4287	4287
R-squared	0.015	0.016	0.007	0.017
Adjusted R-squared	0.012	0.013	0.005	0.013

The second topic is on the research question 2 about whether these two measures can be used complementarily, instead of only as substitutes for each other. My results show support for hypotheses 2a and 2b. When controlling for the value-relevant soft information in earnings calls using the FinancialBERT-sentiment, the Henry's tone measure seems to capture the negative effect of the linguistic tone, managers' strategic use of language in the earnings call's presentation section. The regression models show a clear negative association between the TONE in the presentation section and the cumulative abnormal returns in the initial reaction period (-1, 1), an effect that also persists in the longer term, looking at the full reaction period.

When it comes to the third research question, my study shows a very different phenomenon than expected and hypothesized. If there ever was a post-earnings announcement drift for the soft information contained in the earnings calls (as found by Price et al. 2012), it has disappeared in the same way as the drift for hard information has (Martineau 2021). For the Q&A section of the call, the information seems to be incorporated into share prices immediately within the initial reaction period, with no impact during the extended period. As far as the soft information in the presentation section is concerned, I find a strong initial reaction to the "value-relevant" sentiment, but then a complete reversal of this effect. While the results of the portfolio-level analysis hinted at the possibility that this reversal would be even larger than the initial reaction, these regression results do show that the long-term abnormal returns are no different from zero in a statistically significant way. This suggests that at least with the linguistic tools used in this study, there would be no useful information in the content of the presentation section of the call, apart from the higher linguistic tone predicting lower returns. There is a sizeable investor overreaction to the presentation section's content, when in fact no reaction at all would be warranted.

How to reconcile these results? To start off, we can see the worth in having a Q&A section in earnings calls: The insignificance of Henry's tone measure hints that the managers are unable to strategically control the tone during this section, leading to a relative unbiasedness of this section. The long-term significance of the FinancialBERT sentiment measure shows that the discussion content in this section is relevant to the firm's valuation. These findings seem rather intuitive. Evidence suggests that managers may try to influence the Q&A session in some ways, such as by picking more optimistic analysts' questions (e.g. Mayew 2008). However, it seems plausible that they would be much less successful in controlling the Q&A session's flow relative to that of the presentation section. It is also plausible that the actual incremental informativeness of the call content is high in this section: The analysts asking the questions during the call are professionals with a relatively good understanding of the firm's business, and eagerness to ask the questions that are most relevant for the firm's valuation, which obviously means seeking

information that hasn't been disclosed yet by the firm in e.g. earnings press releases or the call's presentation section.

The results are somewhat more mixed when assessing the tone and sentiment of the presentation section. My findings on the negative impact of the Henry's tone in the initial period is consistent with e.g. Blau et al. (2015) and Baginski et al. (2018), showing that overly positive tone by the management should be considered as bad news. My method of extracting this abnormal tone information from text is however fundamentally different than previous approaches. Previous literature measures abnormality of tone either relative to fundamental and other quantitative information, which does not allow for any value-relevant incremental information to exist in the text or require a less biased baseline text (such as the Q&A section tone, as used by Blau et al. 2015). An additional feature of my approach allows for the "abnormal" tone to be measured in any financial text data, with no need for a baseline text, while still permitting the existence of value-relevant information in the text.

As with the sentiment of the Q&A section, the impact of presentation section's tone is incorporated into share prices completely during the initial reaction period, with no delay or a drift. However, if the perceived speed at which financial markets react to the tone and sentiment of the earnings call testify of highly efficient financial markets, another finding in this paper somewhat subverts this notion. The apparent overreaction to the so-called "value-relevant" sentiment of the presentation section looks interesting. It appears as if though the financial markets first react strongly to this signal in earnings calls, and then slowly reverting to a point where this content bears no meaningful information value at all. There are some potential explanations for this effect. As a reminder, FinancialBERT is trained to classify sentences based on whether the information might cause the stock price of the firm to go up or down. While this sentiment score should therefore measure the value-relevant information, unlike the Henry's tone measure, the managers still have the power to impact the information in the earnings calls' presentation section. To some extent, they might cherry-pick the most positive available information without mentioning as much of the negative pieces of information, making even the "value-relevant" measure less relevant for the firm's valuation, as all this information would be a result of a selection bias driven by managers' incentives to present the most positive information about the firm. Another possibility is that the call content in the presentation section is not that informative incrementally; Much of the management presentation is often just reiterating the developments in the financial performance of the company, which is information that might be more accurately proxied – and therefore captured in the regressions – by the earnings surprise measure. This is different to the Q&A section that can be expected to contain previously unknown information, as that section revolves around the managers answering questions that are asked by the analysts. A third possibility is that there really is previously undisclosed value-relevant

information in the presentation section of the earnings calls, but that this information is so contextual and difficult to decipher, that even the FinancialBERT sentiment analysis model is unable to capture the actual information value properly. However, this explanation would still not explain the measured overreaction.

Whatever explanation might hold, the results nevertheless suggest that in addition to the informational element, there is also a predictable and behavioural element in the cumulative abnormal stock returns following a quarterly earnings call: An overreaction to soft information that is irrelevant for the longer-term valuation of the firm. Regarding this finding, there is an interesting parallel to some of the research studying the impact of abnormal tone. As mentioned earlier, there has been a range of research showing how abnormal tone produces positive immediate reaction followed by a reversal during a longer period of time in for example earnings press releases (X. Huang et al. 2014), management forecasts (Baginski et al. 2018), and even in earnings calls (Hennig et al. 2023). However, what my results suggest, is that this reaction pattern may have been misattributed to linguistic tone in existing literature, despite actually being a pattern related to information content in the managers' statements.

It is also useful to remind ourselves that whatever impact can be observed related to the tone or sentiment of the presentation section of the earnings call might in fact be an impact related to the soft information in the earnings press releases. As was noted before, the content in presentation sections of the calls very often align with the content in the earnings press releases. Keeping this in mind, one possible explanation for the overreaction is that there is a sizeable group of less sophisticated noise traders, who cause these overreactions based on the more easily available earnings press releases, instead of the actual earnings calls. To provide some additional insights into potential explanations such as that, in the next section, I study the differences in these uncovered stock return phenomena between firms with different ownership bases and information environments.

6 Additional analysis

To better understand the phenomena behind the regression outputs, I also conduct some supplementary analysis on the results. While I leave most of the in-depth analysis about the varied responses to the studied phenomena between different firms for future research, in this chapter I provide some additional findings. I aim to provide an improved idea of how firms with different characteristics or information environments might be affected by the impact of tone and sentiment in quarterly earnings calls. It is good to note, however, that splitting the sample into smaller subsamples obviously somewhat limits the statistical power of the regressions. As a result, many of the coefficient estimates in this analysis are relatively sizeable, while still not reaching sufficient levels of statistical significance. Furthermore, while the cumulative abnormal returns have relatively good information-to-noise ratio during the initial 3-day stock reaction period (as indicated by e.g. higher R-squared numbers), especially the studies of cumulative abnormal returns over the full reaction period, CAR(-1, 60), are statistically weaker. Indeed, the coefficient estimates rarely reach statistical significance for this period when studying these subsamples of data.

More importantly though, the closer inspection and analysis of potential reasons behind these different responses to the reported phenomena is not attempted here, but rather left for future research. This sample splits are reported with the intention of having a somewhat improved understanding of the functioning of the phenomena that were uncovered in the main analysis. The discussion in this chapter about the potential reasons for the different responses between different firms is offered to the reader as just such: Speculation about where future research agendas might be directed for possible answers to these questions.

Furthermore, one unanswered foundational question is whether these different phenomena arise from informationally efficient markets or behavioural explanations. As an example, one difference noted above was the stronger reactions to Q&A sentiment for firms with higher trading volume than lower trading volume. Based on the knowledge that can be extracted from the analysis below, it is not possible to say whether some of these variances arise from differences in actual information value under efficient markets (Are the Q&A sessions more informative for firms with higher investor attention?), or behavioural factors and information processing (Are investors acting more on the information provided during the calls for firm with higher investor attention?). Therefore, in this section, I have not aimed to provide analysis to test the reasons behind any reported differences, or evidence in favour of one theory over another, but rather provide an overview of the diversity of potential explanations.

6.1 Regressions of firms with different earnings surprises

Table 8 shows the regression results for (i) the initial, (ii) extended, as well as the (iii) full stock price reaction periods for observations where the earnings surprise was negative (left side of the table), and for observations where the earnings surprise was positive (right side of the table). In my sample, there were a total of 1410 observations of the former, and 2582 observations of the latter type.

Table 8. Regression results on the cumulative abnormal returns for observations with positive and negative earnings surprise

This table shows the key regression statistics for some of the relevant independent variables on the cumulative abnormal returns for two subsamples of the data: Observations with a negative earnings surprise (left half of the table) and observations with a positive earnings surprise (right half of the table). The regression results are for three different sample periods' cumulative abnormal returns as the dependent variable: The initial reaction period CAR(-1, 1), the extended reaction period CAR(2, 60), and the full reaction period CAR(-1, 60). "TONE presentation" and "TONE Q&A" are the Henry's tone measures for the earnings calls' presentation and Q&A sections, respectively. Similarly, "SENT presentation" and "SENT Q&A" are the FinancialBERT sentiment measures for the earnings calls' presentation and Q&A sections, respectively. The models that these results show the summaries of are similar to the previously reported models with full control variables. In other words, the other independent variables (not reported in this summary table) in the regressions also include earnings surprise, the word count controls for presentation and Q&A section, as well as the other controls (firm size, book-to-market-equity, profitability, leverage, trading volume, returns volatility, analyst coverage and dividend declaration). The dependent variable CARs are in percentages. T-statistics for the coefficient estimates are in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1

	Negative SURP			Positive SURP		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-8.559***	4.088	-4.471	0.147	-7.272**	-7.124*
	(-4.26)	*(0.67)*	*(-0.66)*	*(0.11)*	*(-2.05)*	*(-1.89)*
TONE Q&A	2.211	-1.905	0.306	0.778	3.238	4.016
	(1.26)	*(-0.38)*	*(0.06)*	*(0.48)*	*(0.75)*	*(0.85)*
SENT pres.	9.885***	-6.966	2.919	2.176	-7.641**	-5.465*
	(5.32)	*(-1.27)*	*(0.47)*	*(1.49)*	*(-2.41)*	*(-1.65)*
SENT Q&A	3.901**	-0.213	3.688	4.348***	1.615	5.963*
	(2.51)	*(-0.05)*	*(0.71)*	*(3.51)*	*(0.55)*	*(1.85)*
constant	3.425	0.294	3.719	-3.338	6.773	3.436
	(1.35)	*(0.04)*	*(0.44)*	*(-1.59)*	*(1.17)*	*(0.55)*
Observations	1410	1410	1410	2582	2582	2582
R-squared	0.118	0.031	0.028	0.047	0.029	0.027
Adj. R-squared	0.108	0.021	0.018	0.042	0.023	0.022

Looking at the results, there are a few notions to be made. Regarding the value-relevant sentiment, the effects of soft information in the Q&A section of the call seem to be qualitatively similar between both types of firms: A higher sentiment is associated with positive cumulative abnormal returns

that materialize in the initial stock reaction period. However, this impact appears to be slightly stronger for firms with a positive earnings surprise: The coefficient estimate for SENT Q&A is 4.3 for firms with a positive earnings surprise (significant at 1-% level), whereas the same coefficient is 3.9 for firms with a negative earnings surprise (statistically significant at 5-% level). Accordingly, only the Q&A sentiment for the positive earnings surprise firms has a statistically significant (albeit at only 10-% level) impact during the full reaction period.

Another difference can be found in the reaction to the sentiment of the presentation section. As I previously reported, the aggregate reaction to the presentation section sentiment can be described as an overreaction, with significantly positive association with the abnormal returns during the initial period and a negative association in the extended period. The results in Table 8 suggest an alternative explanation: The initial positive association is focused strongly to firms with a negative earnings surprise, whereas the negative effect during the extended period is ever so slightly stronger for firms with a positive earnings surprise. The coefficient estimate for the initial impact of presentation sentiment is 9.9, significant at 1-% level, for observations with a negative earnings surprise, which is a much stronger reaction than the result obtained for the full sample. Alternatively, the coefficient estimate for observations with a positive earnings surprise is a statistically insignificant 2.2. However, the slight difference in the reversal during the extended period is much more subtle: The reversal during the extended period is estimated as almost equal in size for both groups: -7.0 for negative earnings surprise firms and -7.6 for positive earnings surprise firms, although only the latter is statistically significant at 5-% level. While it might not be the full explanation, this does exhibit the issue of somewhat reduced statistical power that is made as a sacrifice when analysing a part of the full sample in this manner.

Finally, while the abnormal tone of the presentation section, as measured with Henry's tone, is negatively associated with the cumulative abnormal returns for both firms, the timing of this impact varies. For firms with a negative earnings surprise, the impact is immediate: The effect is of a considerable size as well as statistically significant on 1-% level during the initial stock reaction period, but insignificant in the extended period. For observations with positive earnings surprise, the opposite is true: the negative impact of abnormal tone seems to become incorporated into stock prices with a delay, only during the extended reaction period. Keeping in mind that average earnings call presentations have a positive tone, it is possible that investors are initially more lenient towards those managers' inflated tone who present positive earnings news, and conversely harsher and more sceptical towards managers who present negative earnings news, leading to a delayed reaction to the former type of firms.

However, the conclusions that can be drawn from these results are somewhat limited. As mentioned, filtering into these two smaller subsets from the full sample reduces the statistical power of these tests, and there is also some low-to-moderate multicollinearity in the regressions. Both issues are stronger especially with the subsample of observations with negative earnings surprise. This can also be seen from some of the coefficient estimates that despite being quite sizeable, are ultimately just not statistically significant.

6.2 Regressions of firms with different firm characteristics

Finally, parallel to the results for observations with either a negative or positive earnings surprise, I report key statistics for similar regressions done for subsamples based on five other factors of firm characteristics. Table 9 presents these result summaries in the following manner: Panel A shows the results when the full sample is divided into two halves based on market capitalization, into smaller and larger firms, and Panel B reports the results for when the sample is divided based on the firm book-to-market equity, into low B/M and high B/M halves. Accordingly, the remaining data split summaries are presented in Panel C (low and high trading volume), Panel D (low and high share price volatility), and Panel E (low and high share of institutional ownership, for the part of the sample where this ownership data is available). It is again good to note, that the same statistical reservations noted previously apply here as well; The at times sizeable coefficient estimates combined with somewhat weakened statistical significance hint at the slightly less generalizable results than in the main analysis in Chapter 5. However, that does not mean that there wouldn't be any conclusions to be drawn from these results. As a remainder, there are three key phenomena uncovered in the main analysis, that are of the greatest interest here: (i) The negative impact of presentation section's linguistic tone on cumulative abnormal returns ("TONE presentation"), (ii) the apparent overreaction, and a subsequent reversal on the value-relevant information in the presentation section ("SENT presentation"), and finally (iii) the positive impact of value-relevant information of the Q&A section of the call ("SENT Q&A").

Table 9. Regression results on the cumulative abnormal returns for observations with different firm characteristics

This table shows the key regression statistics for some of the relevant independent variables on the cumulative abnormal returns for different splits of the data. For these regressions, the data has been divided into two halves based on firm size (Panel A), book-to-market Equity (Panel B), trading volume (Panel C), volatility (Panel D), and the share of institutional ownership (Panel E). The regression results are for three different sample periods' cumulative abnormal returns as the dependent variable: The initial reaction period CAR(-1, 1), the extended reaction period CAR(2, 60), and the full reaction period CAR(-1, 60). "TONE pres." and "TONE Q&A" are the Henry's tone measures for the earnings calls' presentation and Q&A sections, respectively. Similarly, "SENT pres." and "SENT Q&A" are the FinancialBERT sentiment measures for the earnings calls' presentation and Q&A sections, respectively. The models that these results show the summaries of are similar to the previously reported models with full control variables. In other words, the other independent variables (not reported in this summary table) in the regressions also include earnings surprise, the word count controls for presentation and Q&A section, as well as the other controls (firm size, book-to-market-equity, profitability, leverage, trading volume, returns volatility, analyst coverage and dividend declaration). The dependent variable CARs are in percentages. T-statistics for the coefficient estimates are in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1

Panel A: Size

	Small			Large		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-3.300*	-6.054	-9.354	-2.669**	-2.588	-5.257
	(-1.67)	*(-1.08)*	*(-1.54)*	*(-2.07)*	*(-0.85)*	*(-1.56)*
TONE Q&A	1.016	3.435	4.451	1.472	-0.878	0.595
	(0.58)	*(0.72)*	*(0.88)*	*(1.02)*	*(-0.22)*	*(0.14)*
SENT pres.	8.371***	-7.017	1.355	4.568***	-6.240**	-1.673
	(4.07)	*(-1.32)*	*(0.23)*	*(3.67)*	*(-2.26)*	*(-0.58)*
SENT Q&A	5.166***	2.557	7.722*	4.183***	-0.810	3.373
	(3.63)	*(0.73)*	*(1.96)*	*(3.96)*	*(-0.28)*	*(1.08)*
constant	-1.973	4.440	2.467	-0.957	-4.007	-4.965
	(-0.73)	*(0.55)*	*(0.29)*	*(-0.47)*	*(-0.76)*	*(-0.85)*
Observations	2143	2143	2143	2144	2144	2144
R-squared	0.085	0.022	0.020	0.055	0.042	0.024
Adj. R-squared	0.079	0.015	0.013	0.048	0.036	0.017

Panel B: B/M Equity

	Low			High		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-1.893	-4.620	-6.513	-2.524*	-3.748	-6.271
	(-1.09)	*(-1.04)*	*(-1.31)*	*(-1.66)*	*(-0.87)*	*(-1.39)*
TONE Q&A	0.136	-1.473	-1.337	1.667	3.584	5.251
	(0.07)	*(-0.30)*	*(-0.25)*	*(1.14)*	*(0.84)*	*(1.15)*
SENT pres.	6.677***	-6.489*	0.188	5.297***	-7.242*	-1.945
	(3.88)	*(-1.75)*	*(0.04)*	*(3.64)*	*(-1.74)*	*(-0.44)*
SENT Q&A	6.649***	1.370	8.019**	3.227***	1.217	4.445
	(4.52)	*(0.39)*	*(2.09)*	*(2.63)*	*(0.36)*	*(1.19)*
constant	-4.020	4.465	0.445	-0.964	1.337	0.372
	(-1.53)	*(0.65)*	*(0.06)*	*(-0.49)*	*(0.21)*	*(0.06)*
Observations	2143	2143	2143	2144	2144	2144
R-squared	0.087	0.019	0.019	0.076	0.029	0.018
Adj. R-squared	0.081	0.012	0.012	0.069	0.022	0.011

Table 9 (continued)

Panel C: Trading Volume

	Low			High		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-1.451	-9.386**	-10.837**	-4.162**	-0.015	-4.177
	(-0.93)	*(-1.98)*	*(-2.08)*	*(-2.45)*	*(0.00)*	*(-0.99)*
TONE Q&A	-0.195	7.662*	7.466	2.562	-6.523	-3.961
	(-0.13)	*(1.77)*	*(1.61)*	*(1.28)*	*(-1.35)*	*(-0.74)*
SENT pres.	6.456***	-2.996	3.461	5.707***	-8.773***	-3.066
	(4.41)	*(-0.68)*	*(0.71)*	*(3.46)*	*(-2.65)*	*(-0.85)*
SENT Q&A	3.237***	0.605	3.841	6.940***	1.239	8.179**
	(2.74)	*(0.18)*	*(1.04)*	*(4.45)*	*(0.37)*	*(2.21)*
constant	0.098	17.859**	17.957**	2.606	2.189	4.795
	(0.04)	*(2.13)*	*(2.00)*	*(0.66)*	*(0.25)*	*(0.49)*
Observations	2143	2143	2143	2144	2144	2144
R-squared	0.072	0.026	0.022	0.083	0.040	0.028
Adj. R-squared	0.065	0.019	0.015	0.077	0.033	0.021

Panel D: Volatility

	Low			High		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-0.601	-1.650	-2.251	-4.571**	-6.231	-10.801**
	(-0.48)	*(-0.63)*	*(-0.80)*	*(-2.56)*	*(-1.27)*	*(-2.01)*
TONE Q&A	0.519	-0.606	-0.087	1.566	2.931	4.497
	(0.37)	*(-0.21)*	*(-0.03)*	*(0.90)*	*(0.58)*	*(0.84)*
SENT pres.	5.239***	-4.117	1.122	6.866***	-9.101**	-2.236
	(4.42)	*(-1.64)*	*(0.44)*	*(4.04)*	*(-2.20)*	*(-0.48)*
SENT Q&A	5.086***	2.470	7.556***	4.560***	0.055	4.615
	(4.67)	*(1.03)*	*(2.80)*	*(3.22)*	*(0.01)*	*(1.10)*
constant	0.956	2.406	3.361	-3.016	3.191	0.175
	(0.52)	*(0.62)*	*(0.76)*	*(-1.18)*	*(0.40)*	*(0.02)*
Observations	2143	2143	2143	2144	2144	2144
R-squared	0.113	0.018	0.026	0.063	0.029	0.020
Adj. R-squared	0.107	0.011	0.019	0.056	0.022	0.013

Panel E: Share of Institutional Ownership

	Low			High		
	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)	CAR(-1, 1)	CAR(2, 60)	CAR(-1, 60)
TONE pres.	-1.031	-4.635	-5.666	-4.450**	1.165	-3.285
	(-0.65)	*(-0.98)*	*(-1.10)*	*(-2.38)*	*(0.29)*	*(-0.74)*
TONE Q&A	0.425	1.971	2.395	1.747	-0.248	1.499
	(0.24)	*(0.39)*	*(0.45)*	*(0.84)*	*(-0.05)*	*(0.28)*
SENT pres.	4.777***	-7.890**	-3.113	8.271***	-13.549***	-5.278
	(3.62)	*(-2.04)*	*(-0.74)*	*(4.53)*	*(-3.81)*	*(-1.36)*
SENT Q&A	5.046***	1.006	6.052	4.586***	5.607*	10.193***
	(3.59)	*(0.26)*	*(1.45)*	*(3.15)*	*(1.70)*	*(2.80)*
constant	-2.331	4.824	2.494	-0.518	-6.320	-6.838
	(-1.07)	*(0.70)*	*(0.35)*	*(-0.18)*	*(-0.96)*	*(-0.92)*
Observations	1914	1914	1914	1914	1914	1914
R-squared	0.073	0.030	0.029	0.073	0.044	0.024
Adj. R-squared	0.066	0.023	0.021	0.066	0.037	0.016

6.2.1 The negative impact of presentation tone

On the negative impact of the linguistic tone in the presentation section, there are a few observations to be made from these results. Firstly, this phenomenon seems to behave very differently between firms with different levels of trading volume. For observations in the lower half of the sample, the results show that the long-term negative impact is largest of all different data splits, with the coefficient of -10.8 for the full reaction period (statistically significant at 5-% level). It is the timing of this impact, that also stands out: The initial reaction period impact is not significantly different from zero, but the effect seems to be incorporated into share prices almost fully during the extended reaction period, where the coefficient is -9.4, likewise significant at 5-% level. For firms in the upper half of the sample in terms of trading volume, this effect is much more muted, but instantly occurring during the initial reaction period, with a coefficient of -4.2. There are a few possible ways to explain these results. Trading volume has been often connected to the level of investor attention (e.g. Hou et al. 2009). Hirshleifer et al. (2009) test their hypothesis that limited investor attention causes market underreactions. The authors find that the immediate price reaction is weaker, but the post-earnings announcement drift is stronger for these types of firms. My results would therefore seem to hint at the idea that these effects might also be true for the soft information in the managers' tone, with observations with higher investor attention having a more immediate reaction, and observations with lower investor attention a delayed response. Although, the question on why the overall impact of tone appears to be so much larger for low trading volume firms, would still not have an obvious answer.

A related idea to investor attention is the one of investor disagreement. For example, Hong & Stein (2007) discuss the idea of disagreement models, as well as the connection between trading volume and stock prices, with a key idea behind these models often being that investors' disagreement might generate increased trading activity, and therefore lead to higher trading volumes. The authors note that there appears to be a positive association between higher prices (relative to fundamental values) and higher trading volumes, of which asset pricing bubbles are just an extreme example of. This idea is related to Miller's (1977) idea that stock prices tend to reflect optimists' views. This phenomenon is thought to arise in the presence of investor disagreement and short-selling constraints and is supported by a range of studies (e.g. Diether et al. 2002; Daniel et al. 2023). The much lower negative impact of management presentation tone could very well be a product of a disagreement-borne phenomenon: It is very plausible to imagine that a higher, more optimistic linguistic tone would be considered by some as a good sign, despite the evidence hinting at the opposite. This could very well lead to a muted reaction to the tone, driven by the optimists' views for firms with a generally higher investor disagreement, as proxied by the trading

volume. Indeed, as mentioned before, for example Blau et al. (2015) do indeed find evidence of investors interpreting the abnormal tone in earnings calls differently.

Another notion that can be drawn from the regression results of the presentation tone, is that the negative impact of tone is considerably large and immediate for firms with a higher share price volatility, with a coefficient estimate of -4.6 during the initial reaction period, and -10.8 for the full reaction period (both significant at 5-% level). Conversely, there is no statistically significant impact for low volatility firms during any reaction period. A few explanations could be offered for this phenomenon. In general, more volatile shares might react more strongly to new information in general, leading to pronounced changes in share prices. Higher volatility might be a sign of a weaker information environment around the firm, making any reactions to new information more sizeable than firms with a lower volatility. While a relatively speculative statement, this and some of the previous ideas discussed in this section would fit together with the final notion on the reaction to presentation section tone, which is that the reaction is especially during the initial reaction period more significant both statistically and economically for firms with a higher share of institutional ownership. Observations with a higher institutional ownership have a presentation tone coefficient estimate at -4.5 with a 5-% level significance, as opposed to non-significant coefficient estimates for observations with low institutional ownership. As institutional investors are often seen as more sophisticated ones, their higher ownership share is often seen as a sign of a higher informational efficiency (as shown by e.g. Boehmer & Kelley 2009), which would explain the immediate reaction to the negative information inherent in the positive tone. Moreover, as Blau et al. (2015) suggest in their study, sophisticated investors are more likely to accurately assess a higher (abnormal) linguistic tone as bad news.

6.2.2 The overreaction and reversal on presentation sentiment

Moving from the linguistic tone onto financial market's reactions to the so-called value-relevant information, FinancialBERT sentiment of the presentation section, there are also some notions worth highlighting. Something that unites all the different subsamples is that the initial positive reaction to the presentation sentiment is present and statistically significant at 1-% level. Accordingly, the full reaction period impact is insignificant for all subsamples. Thus, the long-term relevance of the presentation sentiment appears indeed questionable across the board.

Despite these similarities, some differences do arise from the data splits. Firstly, while the reversal during the extended period is statistically insignificant for most subsamples, it is highly significant for firms with high past trading volume. It is not immediately apparent how this result can be reconciled with the prior interpretation of these phenomena. As mentioned,

trading volume has been used to proxy for either the level of investor attention or investor disagreement. If higher investor disagreement tends to lead to the optimists' views dominating the price discovery process, this very sizeable reversal is an opposite reaction than one might expect. However, a higher investor attention might offer a more plausible explanation: If we accept as a fact that the presentation sentiment is indeed not very informative in the long-run, higher investor attention might help in correcting the initial overreaction more completely, than for firms with lower investor attention.

Secondly, the initial reaction and the subsequent reversal appear stronger for firms with higher past share price volatility. Here the previously discussed idea that share price volatility would proxy for overall informational efficiency could be used to explain this result. The initial overreaction as well as the following reversal could both be more pronounced for firms due to a lower informational efficiency (i.e. firms with higher volatility) leading to a more erratic price discovery process.

Thirdly, both the initial reaction as well as the following reversal appear to be stronger for firms with a higher institutional ownership; The coefficient estimate for these firms for the initial reaction period is 8.2, and for the extended reaction period -13.5 (both significant at 1-% level), in comparison to the coefficient estimates of 4.8 and -7.9 respectively for firms with lower institutional ownership (significant at 1-% and 5-% levels respectively). This is an intriguing result, if we assume that higher institutional ownership acts as a proxy for investor sophistication and thus a possible sign for higher informational efficiency. If anything, these results suggest the opposite, a stronger share price overreaction for these types of firms. One possible explanation could be the impact of institutional herding. While higher share of institutional ownership is found to be associated with higher informational efficiency (e.g. Boehmer & Kelley 2009), Nofsinger & Sias's (1999) findings suggest that institutional investors might either conduct more positive-feedback trading than individual investors, or that the herding behavior by institutional investors affects stock prices more than herding by individual investors.

6.2.3 The impact of Q&A section's sentiment

Finally, there is the phenomenon of positive impact of value-relevant information in the earnings call's Q&A section, as shown by variable "SENT Q&A". Unsurprisingly, this effect is statistically highly significant during the initial reaction period for every subsample. Furthermore, the effect is very consistently of a roughly equal size for each subsample, with much less variation between different firms than with the previous two phenomena. However, just like before, the statistical significance does suffer when looking at the full reaction period. The positive impact of Q&A section sentiment seems to be the clearest and most persistent for four types of observations: firms with

either (i) high institutional ownership, (ii) high trading volume, (iii) low book-to-market equity, or (iv) low volatility.

For firms with high institutional ownership share, the effect is by far largest in the full reaction period, with a coefficient estimate of 10.2, which is significant at 1-% level. There are a few reasons why that might be. As already mentioned, institutional herding might affect share prices more than individual investors' herding, which might contribute to the effect. However, a possibly even more likely explanation is that it is the more sophisticated investors (such as institutional investors) that follow earnings calls more closely than individual investors. As a remainder, the effects stemming from the presentation section of the call might be capturing impact from the earnings press releases, due to similar contents, whereas Q&A section is much more robustly measuring the information provided specifically during the earnings call. As individual investors might attend these calls less often, the share prices for firms with higher individual investor ownership might not therefore react to the call content as strongly.

As mentioned above, trading volume can act as a proxy for either investor attention or investor disagreement, either of which could be used to explain the larger reaction to earnings call content: Higher investor attention could mean that earnings calls for those firms are also observed more closely, leading to more trading based on that soft information.

It is similarly not immediately evident why firms with low book-to-market equity would have a stronger reaction the Q&A section sentiment. One possible explanation is that the valuation of growth firms (with low B/M equity) is driven more by expectations of future performance, instead of the current financial performance, making the types of soft information more relevant to growth firms than for value firms. Of course, it is also possible that there are reactions that are for different reasons varying between the two types of firms. For example, Skinner & Sloan (2002) document an asymmetric reaction to earnings news between growth and value stocks, showing that growth stocks react more strongly to negative earnings surprises.

Finally, the impact of Q&A sentiment appears more robust for firms with a lower volatility than for firms with higher volatility. This seems initially at least somewhat counterintuitive, as it could be expected that firms with higher volatility are associated with more pronounced price movements. Of course, the question could also be framed in another way: Why is the reaction weaker for firms with high volatility? It is possible that firms with higher informationally efficient stock prices (i.e. firms with lower volatility), the soft information in earnings calls' Q&A section is incorporated into share prices more fully.

7 Conclusion

Price formation of stock prices around firms' information events has not always adhered to the expectations of informatively efficient markets. Instead, these information events, such as earnings announcements and other disclosure releases have been surrounded by return phenomena consistent with behavioural anomalies such as over- or underreaction, or in other ways slow incorporation of new information into share prices. It is also known that many of these phenomena have largely disappeared, and market do seem to have become more informationally efficient in the recent past. In this thesis, I have studied the price formation of stocks surrounding and following the release of soft information in quarterly earnings calls. Specifically, I have studied stock price reactions to the sentiment and tone of these earnings calls. Building on the work on LLM-based sentiment analysis tool by Hazourli (2022b), I test and demonstrate the viability and usefulness of this previously underutilised approach in financial textual content analysis. Additionally, I provide a specific theoretical distinction between the call content's tone and sentiment, showing how researchers can use these novel tools to study the exact phenomena and different content characteristics more accurately, instead of being bound by the limitations of previous methods.

There are three key findings that I provide in this thesis. Initially, I show how the LLM-based FinancialBERT can be used to gauge the sentiment of firms' quarterly earnings calls, meaningfully outperforming the previous methodology. I then demonstrate how using both FinancialBERT and Henry's (2008) dictionary -approach together allows us to isolate different aspects of the call content: The value-relevant information content, as well as the linguistic tone. Utilising the benefits provided by these contributions, I am able to identify three distinct phenomena related to how soft information in quarterly earnings calls affect stock returns. Firstly, I show that while there is incrementally value-relevant information content in the Q&A sections of the earnings calls, the stock price drift documented by Price et al. (2012) does not exist anymore. Secondly, I show that when controlling for value-relevant information content, a higher managers' linguistic tone during the calls' presentation section predicts lower abnormal returns. Furthermore, how quickly this negative impact is incorporated into share prices varies between different firms: The impact is much faster for firms that presented a negative earnings surprise, contrary to the slower reaction for firms with a positive earnings surprise. Finally, I show that the seemingly value-relevant information content in a call's presentation section leads to a type of an overreaction and a subsequent reversal on this information. This reversal is so large, that it renders the long-term value-relevance of this information non-significant.

Some academics have voiced concerns about the use of simple dictionary-based methods as a norm in studies employing computational linguistics

in finance and accounting disciplines (e.g. El-Haj et al. 2019). Unlike the existing literature, I study the information in earnings calls utilizing FinancialBERT. This is a sentiment analysis tool based on the BERT large language model, that has been further trained with a large set of finance-specific text data and fine-tuned with a hand-annotated Financial PhraseBank dataset designed for financial sentiment analysis. This approach not only provides a superior performance to previous methods, but it also opens completely new avenues for research, such as separating the impact of information content and linguistic tone in text.

In the broader context, the field of textual sentiment analysis in finance has suffered from rudimentary methodology set, but also from another problem: the lack of a mutually agreed and understood theoretical definition of tone or sentiment. This lack of ambition in defining these core terms has only few exceptions (e.g. Henry 2008; Tetlock 2007). While different definitions have been proposed, none of them have truly caught on in the studies that attempt to study the impact of tone or sentiment. Ultimately, researchers have also been bound by the limitations of the existing methodologies, where even a better theoretical understanding of the tone or sentiment as a textual feature would not have helped much. This is an issue that my use of both FinancialBERT and Henry's dictionary helps address.

Additionally, my finding that there is an overreaction and a subsequent reversal on the value-relevant textual sentiment, not the linguistic tone, provides a new perspective on how to understand the phenomenon that has been discussed in previous literature. Furthermore, my finding of the actual value-relevant information in earnings calls' Q&A sections shows a clear deviation from the previous literature's findings about the slow incorporation of earnings calls' soft information into share prices. My results provide in many ways a completely new narrative about what kind of soft information is useful for investors, and how that information is incorporated into share prices.

Nevertheless, there are some limitations to this study. The sample, while designed to be as representative of the overall market as possible, might be skewed in certain ways: Firstly, I have limited my sample into firms with at least some analyst following to control for the analyst earnings surprise. While this is the often-recommended way to study earnings surprise related phenomena (Martineau 2021), it does introduce limitations to the sample selection process. Furthermore, the sample naturally only consists of observations for which the quarterly earnings call was available. Additionally, this study cannot reliably distinguish between the impact of soft information in the earnings call's management presentation section and earnings press release. As Price et al. (2012) note, managers often use the presentation section to just simply reiterate what has been disclosed in the earnings press release. Moreover, while the FinancialBERT tool is outperforming the previous method, it is not perfect. Despite its high accuracy, it can still make errors, especially if there is important context in other sentences than the one being

evaluated, as the tool only considers individual sentences separately. Finally, while I have conducted additional analysis on the market reactions to tone and sentiment information for observations with different firm characteristics or information environment, these studies' statistical power does suffer from smaller sample sizes.

There are several attractive avenues for future research in financial textual sentiment analysis. First and foremostly, the more in-depth understanding of the factors contributing to the negative response to management presentation tone, as well as to the overreaction and reversal to the value-relevant information in the same text are fundamental in studying any aspects of earnings calls' soft information. What exactly drives the overreaction is therefore a key question yet to be answered. Another avenue for future research is the further development of methodology. The use of more sophisticated tools in this research is hopefully only just the beginning, and future research should find ways to improve on it. While FinancialBERT's ability to understand contextual clues within a sentence is a major leap to the previous methods, future tools should be developed keeping in mind that there even more value in understanding the context of a sentence as a part of the broader textual content. Secondly, despite the Henry's tone managing to capture the desired phenomenon in this study, it is not a perfect measure for the linguistic tone. Even for this task, researchers could consider the benefits of using a more sophisticated method, such as an LLM-based deep learning model. Finally, the separation of abnormal tone and value-relevant sentiment allows for a much deeper understanding of earnings calls; Researchers could try to better understand the interplay between the two features. Not only that, but my approach for separating the two allows for the abnormal tone to be identified in any financial disclosure: While previous methods of measuring abnormal tone in financial text have most often relied on studying the difference between the management presentation section and the Q&A section (Blau et al. 2015) or the difference between the overall tone of a disclosure and some set of quantitative variables (e.g. Baginski et al. 2018; X. Huang et al. 2014), with the inclusion of separate measures for abnormal tone and sentiment in the same text allows researchers to study the abnormal tone of any financial disclosure, such as annual reports or firm press releases.

Anybody can understand the power and importance of the narrative for firm's valuation and standing in the financial markets. Developments in content analysis methodologies for finance, as well as improving our understanding of the texts and narratives and their impact in the financial is still a somewhat under-researched field. Evolving our capabilities in understanding soft information in financial textual data will therefore most likely have a lot more to tell us about the inner workings of financial markets.

References

Abarbanell, J. S., & Bernard, V. L. (1992). Tests of Analysts Overreaction Underreaction to Earnings Information as an Explanation for Anomalous Stock-Price Behavior. Journal of Finance, 47(3), 1181-1207. https://doi.org/10.2307/2328982

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics Meets Sentiment: An Overview of Methodology and Applications. Journal of Economic Surveys, 34(3), 512-547. https://doi.org/10.1111/joes.12370

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

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Athanasakou, V., & Hussainey, K. (2014). The perceived credibility of forward-looking performance disclosures. Accounting and Business Research, 44(3), 227-259. https://doi.org/10.1080/00014788.2013.867403

Baginski, S. P., Demers, E., Kausar, A., & Yu, Y. J. (2018). Linguistic tone and the small trader. Accounting Organizations and Society, 68-69, 21-37. https://doi.org/10.1016/j.aos.2018.03.005

Baginski, S. P., Hassell, J. M., & Kimbrough, M. D. (2004). Why do managers explain their earnings forecasts? Journal of Accounting Research, 42(1), 1-29.

Ball, R., & Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. Journal of Accounting Research, 6(2), 159-178. https://doi.org/https://doi.org/10.2307/2490232

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep Learning for AI. Communications of the Acm, 64(7), 58-65. https://doi.org/10.1145/3448250

Bernard, V. L., & Thomas, J. K. (1989). Post-Earnings-Announcement Drift - Delayed Price Response or Risk Premium. Journal of Accounting Research, 27, 1-36. https://doi.org/10.2307/2491062

Bhushan, R. (1994). An informational efficiency perspective on the post-earnings announcement drift. Journal of accounting & economics, 18(1), 45-65. https://doi.org/10.1016/0165-4101(94)90018-3

Blau, B. M., DeLisle, J. R., & Price, S. M. (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

Bochkay, K., Hales, J., & Chava, S. (2020). Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls. Accounting Review, 95(2), 31-60. https://doi.org/10.2308/accr-52507

Boehmer, E., & Kelley, E. K. (2009). Institutional Investors and the Informational Efficiency of Prices. Review of Financial Studies, 22(9), 3563-3594. https://doi.org/10.1093/rfs/hhp028

Borochin, P. A., Cicon, J. E., DeLisle, R. J., & Price, S. M. (2018). 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

Brau, J. C., Cicon, J., & McQueen, G. (2016). Soft Strategic Information and IPO Underpricing. Journal of Behavioral Finance, 17(1), 1-17. https://doi.org/10.1080/15427560.2016.1133619

Brockman, P., Li, X., & Price, S. M. (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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In. Ithaca: Cornell University Library, arXiv.org.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-m., & Steele, L. B. (2013). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19(1), 396-455. https://doi.org/10.1007/s11142-013-9258-3

Chen, H. L., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. Review of Financial Studies, 27(5), 1367-1403. https://doi.org/10.1093/rfs/hhu001

Chordia, T., Goyal, A., Sadka, G., Sadka, R., & Shivakumar, L. (2009). Liquidity and the Post-Earnings-Announcement Drift. Financial Analysts Journal, 65(4), 18-32. https://doi.org/10.2469/faj.v65.n4.3

Chordia, T., Subrahmanyam, A., & Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? Journal of accounting & economics, 58(1), 41-58. https://doi.org/10.1016/j.jacceco.2014.06.001

Daniel, K., Klos, A., & Rottke, S. (2023). The Dynamics of Disagreement. Review of Financial Studies, 36(6), 2431-2467. https://doi.org/10.1093/rfs/hhac075

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), 1375-1388. https://doi.org/10.1287/mnsc.1070.0704

D'Augusta, C., & DeAngelis, M. D. (2020). Does Accounting Conservatism Discipline Qualitative Disclosure? Evidence From Tone Management in the MD&A. Contemporary Accounting Research, 37(4), 2287-2318. https://doi.org/10.1111/1911-3846.12598

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-+. https://doi.org/10.1111/j.1911-3846.2011.01130.x

Davis, A. K., & Tama-Sweet, I. (2012). Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A. Contemporary Accounting Research, 29(3), 804-+. https://doi.org/10.1111/j.1911-3846.2011.01125.x

De Amicis, C., Falconieri, S., & Tastan, M. (2021). Sentiment analysis and gender differences in earnings conference calls. Journal of Corporate Finance, 71, Article 101809. https://doi.org/10.1016/j.jcorpfin.2020.101809

Dellavigna, S., & Pollet, J. M. (2009). Investor Inattention and Friday Earnings Announcements. Journal of Finance, 64(2), 709-749. https://doi.org/10.1111/j.1540-6261.2009.01447.x

Demers, E., & Vega, C. (2008). Soft Information in Earnings Announcements: News or Noise? International Finance Discussion Papers,

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Li, F. F., & IEEE. (2009, Jun 20-25). ImageNet: A Large-Scale Hierarchical Image Database. IEEE-Computer-Society Conference on Computer Vision and Pattern Recognition Workshops, Miami Beach, FL.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K., & Assoc Computat, L. (2019, Jun 02-07). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Conference of the North-American-Chapter of the Association-for-Computational-Linguistics - Human Language Technologies (NAACL-HLT), Minneapolis, MN.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. Journal of Finance, 57(5), 2113-2141. https://doi.org/10.1111/0022-1082.00490

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings Conference Call Content and Stock Price: The Case of REITs. Journal of Real Estate Finance and Economics, 45(2), 402-434. https://doi.org/10.1007/s11146-010-9266-z

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. https://doi.org/https://doi.org/10.1111/jbfa.12378

Engelberg, J. (2008). Costly Information Processing: Evidence from Earnings Announcements. AFA 2009 San Francisco Meetings Paper. https://doi.org/https://dx.doi.org/10.2139/ssrn.1107998

Engelberg, J. E., Reed, A. V., & Ringgenberg, M. C. (2012). How are shorts informed? Short sellers, news, and information processing. Journal of Financial Economics, 105(2), 260-278. https://doi.org/10.1016/j.jfineco.2012.03.001

Fei, X. Y., Xu, H. K., & Zhang, J. R. (2023). Linguistic attributes and trade credit: Evidence from textual analysis of earnings conference calls. Journal of Corporate Accounting and Finance, 34(1), 119-136. https://doi.org/10.1002/jcaf.22585

Ferris, S. P., Hao, G., & Liao, S. M. (2013). The Effect of Issuer Conservatism on IPO Pricing and Performance. Review of Finance, 17(3), 993-1027. https://doi.org/10.1093/rof/rfs018

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

French, K. R. (2023). U.S. Research Breakpoints Data, ME Breakpoints. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Breakpoints

Fu, X., Wu, X. X., & Zhang, Z. F. (2021). The Information Role of Earnings Conference Call Tone: Evidence from Stock Price Crash Risk. Journal of Business Ethics, 173(3), 643-660. https://doi.org/10.1007/s10551-019-04326-1

García, D. (2013). Sentiment during Recessions. Journal of Finance, 68(3), 1267-1300. https://doi.org/10.1111/jofi.12027

Guay, W., Samuels, D., & Taylor, D. (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. Journal of Accounting and Economics, 62(2-3), 234-269. https://doi.org/10.1016/j.jacceco.2016.09.001

Hanley, K. W., & Hoberg, G. (2012). Litigation risk, strategic disclosure and the underpricing of initial public offerings. Journal of Financial Economics, 103(2), 235-254. https://doi.org/10.1016/j.jfineco.2011.09.006

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-Level Political Risk: Measurement and Effects. Quarterly Journal of Economics, 134(4), 2135-2202. https://doi.org/10.1093/qje/qjz021

Hazourli, A. (2022a). FinancialBERT for Sentiment Analysis (language model). https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis. Accessed 1.12.2023

Hazourli, A. (2022b (preprint)). FinancialBERT - A Pretrained Language Model for Financial Text Mining. https://doi.org/http://dx.doi.org/10.13140/RG.2.2.34032.12803

Hennig, J. C., Firk, S., & Wolff, M. (2023). Credibility Signals from Soft Information: Evidence from Investor Reactions to Tone in Earnings Conference Calls. European Accounting Review. https://doi.org/10.1080/09638180.2023.2244009

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? The Journal of business communication, 45(4), 363-407. https://doi.org/10.1177/0021943608319388

Henry, E., & Leone, A. J. (2016). Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. Accounting Review, 91(1), 153-178. https://doi.org/10.2308/accr-51161

Hirshleifer, D., Lim, S. S., & Teoh, S. H. (2009). Driven to Distraction: Extraneous Events and Underreaction to Earnings News. Journal of Finance, 64(5), 2289-2325. https://doi.org/10.1111/j.1540-6261.2009.01501.x

Hong, H., & Stein, J. C. (2007). Disagreement and the stock market. Journal of Economic Perspectives, 21(2), 109-128. https://doi.org/10.1257/jep.21.2.109

Hope, O.-K., Hu, D., & Lu, H. (2016). The benefits of specific risk-factor disclosures. Review of Accounting Studies, 21(4), 1005-1045. https://doi.org/10.1007/s11142-016-9371-1

Hou, K., Peng, L., & Xiong, W. (2009). A Tale of Two Anomalies: The Implications of Investor Attention for Price and Earnings Momentum. https://doi.org/https://dx.doi.org/10.2139/ssrn.976394

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

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. Accounting Review, 89(6), 2151-2180. https://doi.org/10.2308/accr-50833

Huang, X., Teoh, S. H., & Zhang, Y. L. (2014). Tone Management. Accounting Review, 89(3), 1083-1113. https://doi.org/10.2308/accr-50684

Irani, A. J. (2004). The Effect of Regulation Fair Disclosure on the Relevance of Conference Calls to Financial Analysts. Review of Quantitative Finance and Accounting, 22(1), 15-28. https://doi.org/10.1023/B:REQU.0000006184.02165.c5

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. Journal of Financial Economics, 110(3), 712-729. https://doi.org/10.1016/j.jfineco.2013.08.018

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

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M. L., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, Article 160035. https://doi.org/10.1038/sdata.2016.35

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

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

Kothari, S. P., Li, X., & Short, J. E. (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. Accounting Review, 84(5), 1639-1670. https://doi.org/10.2308/accr.2009.84.5.1639

Kravet, T., & Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. Review of Accounting Studies, 18(4), 1088-1122. https://doi.org/10.1007/s11142-013-9228-9

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the Acm, 60(6), 84-90. https://doi.org/10.1145/3065386

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting Deceptive Discussions in Conference Calls. Journal of Accounting Research, 50(2), 495-540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

Lee, J. (2016). Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts during Earnings Conference Calls. Accounting Review, 91(1), 229-250. https://doi.org/10.2308/accr-51135

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of accounting & economics, 45(2-3), 221-247. https://doi.org/10.1016/j.jacceco.2008.02.003

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

Liu, Z., Huang, D. G., Huang, K. Y., Li, Z., & Zhao, J. (2021, Jan 07-15). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. 29th International Joint Conference on Artificial Intelligence, Electr Network.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. Journal of Accounting and Economics, 63(1), 1-25. https://doi.org/10.1016/j.jacceco.2016.09.002

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance, 66(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. Journal of Finance, 69(4), 1643-1671. https://doi.org/10.1111/jofi.12162

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. Portland, Oregon, USA.

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. https://doi.org/10.1002/asi.23062

Martineau, C. (2021, forthcoming). Rest in Peace Post-Earnings Announcement Drift. Critical Finance Review. https://dx.doi.org/10.2139/ssrn.3111607

Mayew, W. J. (2008). Evidence of management discrimination among analysts during earnings conference calls. Journal of Accounting Research, 46(3), 627-659. https://doi.org/10.1111/j.1475-679X.2008.00285.x

McLean, R. D., & Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? Journal of Finance, 71(1), 5-32. https://doi.org/10.1111/jofi.12365

Mendenhall, R. R. (2004). Arbitrage Risk and Post-Earnings-Announcement Drift. Journal of Business, 77(4), 875-894. https://doi.org/10.1086/422627

Miller, E. M. (1977). Risk, Uncertainty, And Divergence Of Opinion. Journal of Finance, 32(4), 1151-1168. https://doi.org/10.2307/2326520

Mohanram, P. S., & Sunder, S. V. (2006). How has regulation FD affected the operations of financial analysts? Contemporary Accounting Research, 23(2), 491-525. https://doi.org/10.1506/7h81-8j8x-c6rt-uvjp

Muslu, V., Radhakrishnan, S., Subramanyam, K., & Lim, D. (2014). Forward-looking MD&A disclosures and the information environment. Management Science, 61(5), 931-948.

Nofsinger, J. R., & Sias, R. W. (1999). Herding and feedback trading by institutional and individual investors. Journal of Finance, 54(6), 2263-2295. https://doi.org/10.1111/0022-1082.00188

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. https://doi.org/10.1016/j.jbankfin.2011.10.013

Rennekamp, K. (2012). Processing fluency and investors' reactions to disclosure readability. Journal of Accounting Research, 50(5), 1319-1354. https://doi.org/10.1111/j.1475-679X.2012.00460.x

Sadka, R. (2006). Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk. Journal of Financial Economics, 80(2), 309-349. https://doi.org/10.1016/j.jfineco.2005.04.005

Securities and Exchange Commission (SEC). (2000, 21.8.2000). Retrieved 30.5.2023 from https://www.sec.gov/rules/final/33-7881.htm

Schleicher, T., & Walker, M. (2010). Bias in the tone of forward-looking narratives. Accounting and Business Research, 40(4), 371-390. https://doi.org/10.1080/00014788.2010.9995318

Skinner, D. J., & Sloan, R. G. (2002). Earnings Surprises, Growth Expectations, and Stock Returns or Don't Let an Earnings Torpedo Sink Your Portfolio. Review of Accounting Studies, 7(2-3), 289-312.

Sun, C., Qiu, X. P., Xu, Y. G., & Huang, X. J. (2019, Oct 18-20). How to Fine-Tune BERT for Text Classification? Lecture Notes in Artificial Intelligence. 18th China National Conference on Computational Linguistics (CCL), Kunming Univ Sci & Technol, Kunming, PEOPLES R CHINA.

Tan, H. T., Wang, E. Y., & Zhou, B. (2014). When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments. Journal of Accounting Research, 52(1), 273-302. https://doi.org/10.1111/1475-679x.12039

Tan, H.-T., Wang, E. Y., & Zhou, B. (2015). How does readability influence investors' judgments? Consistency of benchmark performance matters The Accounting Review, 90(1), 371-393. https://doi.org/10.2308/accr-50857

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

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

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. Science, 211(4481), 453-458. https://doi.org/10.1126/science.7455683

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. Journal of Business, 59(4), S251-S278. https://doi.org/10.1086/296365

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

Wang, A., Singh, A., Julian, M., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: a multi-task benchmark and analysis platform for natural language understanding ICLR 2019, https://doi.org/10.48550/arXiv.1804.07461

Wang, X. S., Peng, Y. F., Lu, L., Lu, Z. Y., Bagheri, M., Summers, R. M., & Ieee. (2017, Jul 21-26). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), 165-210. https://doi.org/10.1007/s10579-005-7880-9

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., & Schmidt, L. (2022, Jul 17-23). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. Proceedings of Machine Learning Research. 39th International Conference on Machine Learning (ICML), Baltimore, MD.

Yamamoto, R., Kawadai, N., Kurita, M., & Baba, S. (2022). Managements' tone strategies by earnings call transcripts in the global markets. Journal of Asset Management, 23(3), 246-255. https://doi.org/10.1057/s41260-022-00256-2

Yang, Y., Siy Uy, M. C., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. arXiv.org. https://doi.org/10.48550/arxiv.2006.08097

Zhou, G. F. (2018). Measuring Investor Sentiment. In A. W. Lo & R. C. Merton (Eds.), Annual Review of Financial Economics, Vol 10 (Vol. 10, pp. 239-259). https://doi.org/10.1146/annurev-financial-110217-022725

Appendix 1: Henry's wordlists

Table A1. List of positive words in Henry's (2008) dictionary

above	encouraging	highest	rewarded
accomplish	enjoy	improve	rewarding
accomplished	enjoyed	improved	rewards
accomplishes	enjoying	improvement	rise
accomplishing	enjoys	improvements	risen
accomplishment	exceed	improves	rises
accomplishments	exceeded	improving	rising
achieve	exceeding	increase	rose
achieved	exceeds	increased	solid
achievement	excellent	increases	strength
achievements	expand	increasing	strengthen
achieves	expanded	larger	strengthened
achieving	expanding	largest	strengthening
beat	expands	leader	strengthens
beating	expansion	leading	strengths
beats	good	more	strong
best	greater	most	stronger
better	greatest	opportunities	strongest
certain	grew	opportunity	succeed
certainty	grow	pleased	succeeded
definite	growing	positive	succeeding
deliver	grown	positives	succeeds
delivered	grows	progress	success
delivering	growth	progressing	successes
delivers	high	record	successful
encouraged	higher	reward	up

Table A2. List of negative words in Henry's (2008) dictionary

below	disappointing	low	threat
challenge	disappointment	lower	threats
challenged	disappoints	lowest	uncertain
challenges	down	negative	uncertainty
challenging	downturn	negatives	under
decline	drop	obstacle	unfavorable
declined	dropped	obstacles	unsettled
declines	dropping	penalties	weak
declining	drops	penalty	weaken
decrease	fail	risk	weakened

Table A2 (continued)

decreased	failing	risks	weakening
decreases	fails	risky	weakens
decreasing	failure	shrink	weakness
depressed	fall	shrinking	weaknesses
deteriorate	fallen	shrinks	worse
deteriorated	falling	shrunk	worsen
deteriorates	falls	slump	worsening
deteriorating	fell	slumped	worsens
difficult	hurdle	slumping	worst
difficulty	hurdles	slumps	
disappoint	least	smaller	
disappointed	less	smallest	

Appendix 2. Examples of tone and sentiment classifications

Table A3. Examples of sentences and their classifications of Henry's tone and FinancialBERT sentiment

The net sales forecast continues to incorporate an approximate 1% negative impact from the China JV interest sale in May and hence organic net sales before FX impact would essentially be flat. **For as-reported net sales, we expect to be around 4% lower year-over-year, including an approximate 3% FX headwind versus approximately 2% in our prior assumption and including the other impacts just noted.**	Axalta 2019 Q3 SENT: negative TONE: negative
In our post-ad surveillance, 20% of our respondents stated that they've already had the NovaSure procedure, and 54% stated they are considering it. Albeit a small sample set, we think that it's a very healthy response. **All of this is great but without a revenue pull-through, this investment would not generate an appropriate return, and we're realists about this.**	Hologic Inc. 2011 Q3 SENT: negative TONE: neutral
Our view on the construction market has not changed for this year, and we still expect Electrical Raceway to be up 2% to 4% in 2019. **And although we expect strong industrial markets, we are seeing projects move forward slower than we originally anticipated, and we are lightening our view on MP&S volumes.** We now expect MP&S volumes to be up 2% to 4%.	Atkore International 2019 Q1 SENT: negative TONE: positive
Let's first touch on the hedge book as this is a key differentiator for Antero versus our peer group today. I think it's extremely important to look at how the hedge book was actually constructed. We began building this hedge book more than five years ago and have systematically hedged our risk over this time period rather than trying to lock in pricing when it was too late as the future's curve rolled forward eroding the contango. **For us, this strategy eliminated the problem many E&Ps have faced over the past year, which is the reluctance to hedge at such depressed prices for the short term as commodity prices have continued to fall due to the oversupply.** Our strategy has also been to sell forward our undeveloped gas at prices that we know will generate strong returns and that has paid off, too.	Antero Resources 2015 Q4 SENT: neutral TONE: negative
Each quarter I talk about how well we control our wages and benefits, and we continue to look for new ways to improve and become more efficient when it comes to our community employees' wages. **Recently, we started utilizing a third-party software to help ensure time tracking compliance to reduce our overtime costs.** It is early in the implementation of this overtime initiative with 2 regions covering 36 communities currently using it.	Five Star Senior Living 2018 Q1 SENT: neutral TONE: neutral

Table A3 (continued)	
The business fundamentals are good. Our culture, our execution and our daily intensity around the details of this business makes Potbelly a strong Company. **We believe we are a multidimensional Company that can create value in the marketplace several ways: Company growth, franchise growth, North America growth, growth outside of North America.** Our Company growth is self-funded. We have a strong balance sheet.	Potbelly Corporation 2015 Q1 SENT: neutral TONE: positive
We are proud to announce the new Afton Mill expansion is complete. As expected, we are seeing increases in the recovery of both gold and copper. **The project came in both ahead of schedule and under budget.** We thank our project team for their continued solid execution.	New Gold 2015 Q2 SENT: positive TONE: negative
Before I get into our performance, I would like to take a moment to express our gratitude to the health care providers on the front lines of this pandemic. These are unprecedented times and the challenges facing those caregivers have been unexpected and monumental.	STERIS Plc 2020 Q4 SENT: positive TONE: neutral
Implied billings, excluding Veritas, were $1 billion and grew 3% year-over-year on a reported basis and benefited from a tail wind from currency.	Symantec 2016 Q4 SENT: positive TONE: positive

Table A3 presents a few excerpts from the earnings call transcripts, and information about how the specific sentences in bold are classified or understood by FinancialBERT and Henry's wordlist approaches. The FinancialBERT classification states how the sentence in question was classified by the algorithm as either a positive, neutral, or negative sentence. While Henry's bag-of-words-approach does not classify text as sentences, but rather just counts individual words, the column for Henry's wordlist tone states whether the approach found more positive than negative words in the sentence (positive), equal amounts or no matches in either wordlist (neutral), or more negative than positive words (negative). While a bit of the surrounding text is provided here for the reader as additional context, it is good to note that not even FinancialBERT considers any contextual information provided in the surrounding sentences when making its assessment.

While FinancialBERT does indeed seem to capture the sentiment information better than Henry's approach, the excerpts below are just individual examples to help the reader understand better the very different ideas behind the two methodologies. It is also good to note that while FinancialBERT performs quite well overall, it is nowhere near perfect. Out of any possible combination of different classifications by the two methods, it is possible to find many examples where either tool seems to arrive at a more appropriate assignment. The examples in the table should therefore not be understood as necessarily informative as to what types of sentiments one methodology classifies better than the other.