www.joim.com

Journal Of Investment Management, Vol. 21, No. 1, (2023), pp. 81–102

LEVERAGING TEXT MINING TO EXTRACT INSIGHTS FROM EARNINGS CALL TRANSCRIPTS

Andrew Chin[a,*] *and Yuyu Fan*[a,b]

We apply text-mining techniques in earnings call transcripts to extract meaningful features that capture management and investment community signals. Using a corpus of transcripts of earnings calls for global companies from 2010 to 2021, we create fundamentally driven features spanning document attributes, readability, and sentiment on different sections of the transcripts. We test the efficacy of these features in predicting the future stock returns of companies and find that there are opportunities for investors to use these signals in stock selection. Specifically, we find that readability and sentiment-based techniques can enhance an investor's ability to differentiate amongst outperformers and underperformers and these results are robust across market capitalization as well as investment universes (US Large Cap, US Small Cap, World ex-US, and Emerging Markets). We also introduce methods to create more robust sentiment features for active and systematic investors. By analyzing the performance patterns of the various call participants, we find evidence that the analyst questions may contain more information than the executive sections. Finally, we observe that sentiment features derived from context-driven deep learning language models like BERT are promising and may have more efficacy than bag-of-words approaches.

1 Introduction

As active managers have struggled to outperform over the past decade, fees have come under pressure and investors have abandoned active strategies for passive investments. Indeed, existential questions are being asked about the active management industry—can it survive the current headwinds and reverse the course of history?

Ultimately, investors choose active strategies because of their belief these strategies can outperform their benchmarks over full market cycles. Superior performance may result from exceptional research, deep insights into

[a]AllianceBernstein Holding L.P., 1345 Avenue of the Americas, New York, NY 10105, USA.
E-mail: andrew.chin@alliancebernstein.com
[b]E-mail: yuyu.fan@alliancebernstein.com
*Corresponding author

portfolio construction, strong risk management or efficient trading and implementation techniques. Active managers generate their research edge with sharper insights through a variety of methods—many focus on a selection of stocks that are expected to outperform, typically because these stocks adhere to an investment philosophy or a set of attributes (or factors). Others may delve deeply into the specifics of individual companies to uncover unique insights about their business prospects. When executed successfully, active managers can demonstrate their skill and thus burnish the perception that active managers can add value over time.

Enhancing manager skill can occur in a variety of ways but we will focus on a couple of methods: having access to the most pertinent data and having the ability to quickly synthesize and interpret that data. Asset managers are moving to improve their capabilities in both fronts. Portfolio managers and research analysts are scouring all kinds of data to gain an edge into investments—examples include capturing behaviors and trends in “alternative data”. Many asset managers are also exploring new methods to extract insights from the data. Machine learning and artificial intelligence techniques have been borrowed from other industries and applied to asset management. These trends are set to persist for years to come.

Text mining is one example that captures both trends in the asset management industry. While fundamental analysts have always pored over financial statements and parsed through corporate filings, doing this systematically across all companies, rather than a small subset, offers opportunities for additional insights. In addition, systematic text mining using machine learning algorithms may mitigate biases by removing subjective interpretations by human analysts. Thus, text mining enhances both the breadth and depth of investment analysis.

Various text-mining methods have been explored in financial documents (Tetlock, 2007; Engelberg, 2008). The bag-of-words (BoW) approach is a popular technique and simply counts and scores words in a document in various ways. For example, we can measure the length of a document by simply counting the total number of words, and use this metric to gauge the reading level of the document. We can also extract the significant words or phrases from a document by counting the frequency of those words or phrases, or weighing their impact using the inverse of the relative frequencies across other documents within a corpus. This latter approach is commonly referred to as “Term Frequency–Inverse Document Frequency” (TF–IDF).

A more interesting application of BoW is to score words in a document against different types of dictionaries. Mathematically, this approach is implemented using a document-term matrix with scores assigned to each document against certain terms. The included terms are based on an underlying dictionary that defines the words, phrases, or symbols conveying different types of information. For example, a sentiment dictionary may contain terms expressing emotion and attitude. Since sentiment analysis is one of the most important objectives in text mining, we will provide a brief overview of the relevant research in this area.

Researchers have developed different kinds of dictionaries to measure the sentiment of different types of text. For example, the Harvard Psychosociological Dictionary (Harvard IV-4) is commonly used to measure the sentiment of documents. However, researchers have shown that this dictionary may not be appropriate for sentiment analysis on financial documents because the positive and negative entries may have different connotations in a financial context. For example, a word like “debit” may have negative connotations in common usage but it's a common word in financial settings. To overcome this issue, Loughran and McDonald (2011) developed a dictionary based on the annual corporate filings of US public companies (10-Ks) and found that sentiment scores calculated from this dictionary are correlated to subsequent company stock returns. Their dictionary has been updated several times since its initial publication to account for earnings call transcripts as well as to capture recent changes in semantics. While this dictionary has been widely studied across the asset management industry, one shortcoming has been its inability to capture the degree (valence) of sentiment.

Hutto and Gilbert (2015) developed a rule-based method, Valence Aware Dictionary for Sentiment Reasoning (VADER), to measure sentiment as well as the valence of sentiment for the terms in their dictionary. Unfortunately for asset managers, VADER is more appropriate for social media posts and reviews rather than for financial statements. For example, the VADER dictionary includes various emojis, which are common in posts and reviews, but not in financial documents. In addition, Gentzkow *et al.* (2019) showed the limitations of using static dictionaries in various financial prediction problems and advocated for the usage of more advanced machine learning models.

Recent advances in text mining have demonstrated that more sophisticated techniques can outperform the simple BoW approach. Some studies have developed more flexible and dynamic dictionaries incorporating term valence using advanced machine learning models. For example, Garcia *et al.* (2020) used the Multinomial Inverse Regression Model (MNRM) to generate the sentiment of n-grams from earnings call transcripts using actual stock returns as the main input. While the technique yielded interesting empirical insights, some of the resulting n-grams were difficult to interpret and had little economic intuition.

With the development of more advanced algorithms and improvements in computational power, context-driven language models are becoming more popular. Transformers (Vaswani *et al.*, 2017), a class of deep learning algorithms that can encode and decode the semantic and syntactic information of natural languages, have proven to be effective in various common natural language processing (NLP) tasks. One of the most well-known transformers is the Bidirectional Encoder Representations from Transformers (BERT), a language model pre-trained on large amounts of generic text from Wikipedia and books. BERT has wide applications in NLP tasks and outperforms many benchmarks in numerous tests (Devlin *et al.*, 2019). It is used by Google in its search engine for autocompletion tasks and in machine translation and chatbots to help with various tasks, such as placing orders (Rogers *et al.*, 2020).

Araci (2019) studied the application of the BERT model in finance. They fine-tuned the BERT base model (Devlin *et al.*, 2019) with labelled financial sentences to create the finBERT model. Basically, finBERT added a down-stream classification task to BERT and trained it on human-labeled sentences from financial news. Araci found that finBERT achieved around 85% accuracy in financial sentiment classification and outperformed other machine learning models.

Earnings calls are an important source of information for investors and analysts to assess the financial health of companies over particular periods. These calls provide forums for companies to convey important financial and business information to the investment community and the general public. Investors use the calls to gather information about the industry, the company, its products, its competitors and most importantly, its business prospects. By analyzing the transcripts, investors hope to glean information about a company's future earnings and ultimately, its valuation. Many researchers have used earnings call transcripts to create investment signals (Chebonenko *et al.*, 2018; Garcia *et al.*, 2020).

In our research, we leverage both the BoW approach and BERT to generate a variety of features from earnings call transcripts. We use BoW to capture document attributes, readability, and sentiment on different sections of the transcripts. We then extend Araci's approach with some significant improvements to calculate context-driven sentiment on the transcripts. Our findings are consistent with recent papers which explored the use of text-mining techniques within finance. Klevak *et al.* (2019) found that BoW approaches produce economically significant signals above and beyond traditional factors around earnings announcements. Das *et al.* (2021) trained and fine-tuned the RoBERTa model using corporate filings and showed that their techniques outperformed BoW approaches.

Our paper makes the following contributions to the existing literature. We expand the breadth of features generated from earnings call transcripts and compare them in a systematic way. We find that these features are predictive of future performance in large-cap and small-cap companies, and the results extend across different regions. We also combine features from individual sections of the earnings call transcript by speaker type to create more robust signals for stock selection. In addition, we study the semantic tendencies across different call participants to assess their impact on future stock returns. Finally, we test context-driven techniques like BERT and find that performance can be further enhanced compared to the BoW approach.

In the following sections, we provide an overview of the data and explain how we apply NLP techniques to generate features. We then introduce our backtesting method and show the performance of each category of features. We focus on the sentiment features and compare the performance across different approaches which leverage individual sections and combined sections, as well as amongst different speakers within the transcripts. Finally, we show the efficacy of dictionary-based sentiment features and context-driven sentiment features to assess their potential in differentiating between top and bottom performing stocks.

2 Data

We use the transcripts from earnings calls from S&P Global for our analysis. This dataset is popular within the investment community and is viewed as a high-quality and comprehensive repository of the transcripts from earnings calls. For our study, we use data from January 2010 to December 2021 because the company coverage prior to 2010 was not as comprehensive, especially for non-US companies. Our dataset contains approximately 875,000 transcripts covering about 243,000 calls for more than 12,000 companies.

Exhibit 1 shows the number of companies covered in our dataset each year. Since 2010 the coverage is fairly consistent in the US and Canada. Outside of North America, the coverage has increased over time, capturing the growth of non-US company listings in the US.

Exhibit 2 shows the number of earnings calls by year. Since companies generally hold about four calls each year to correspond with their quarterly submissions of financial statements, the number of calls is around 4× the number of companies in Exhibit 1.

Exhibit 1. Total number of companies by region by year based on event time.

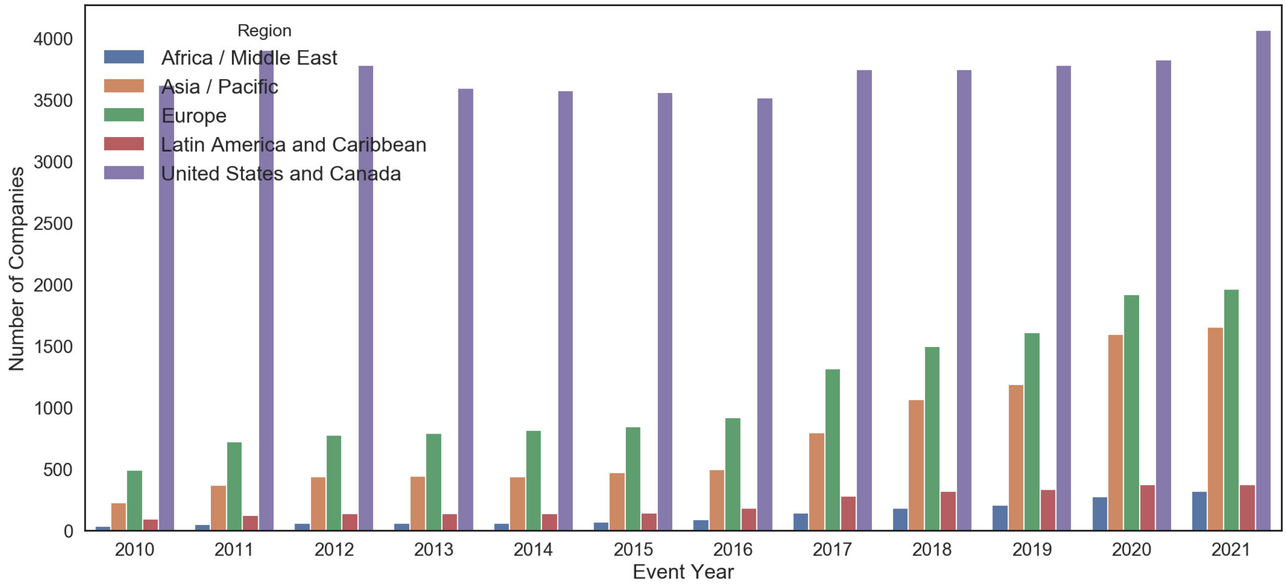


Exhibit 2. Total number of earnings calls by region by year based on event time.

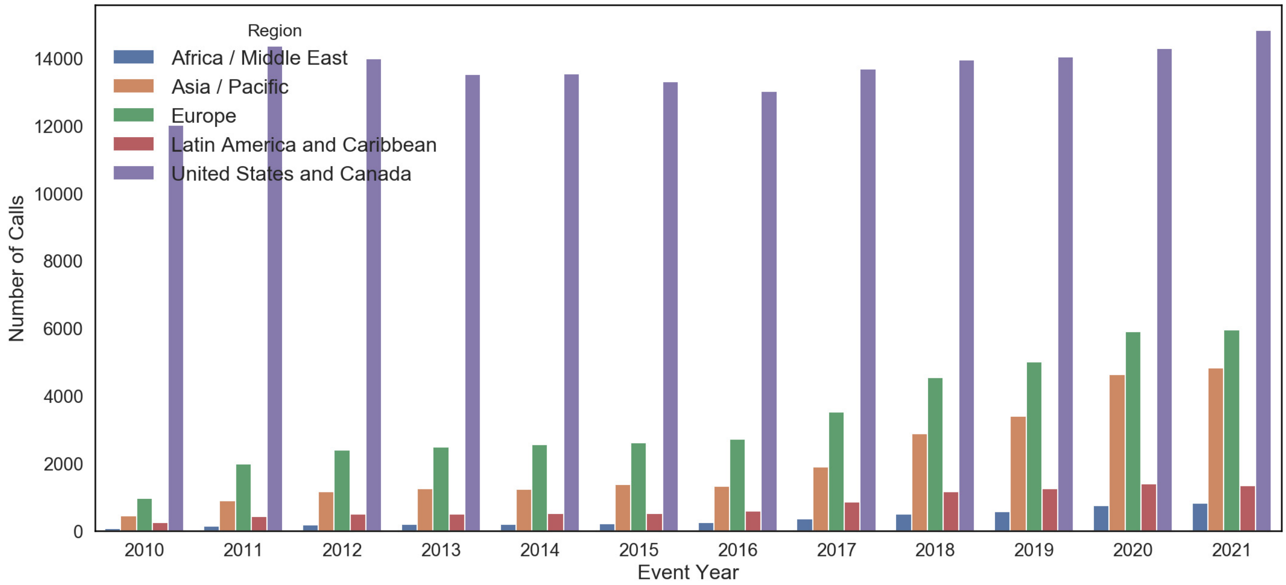
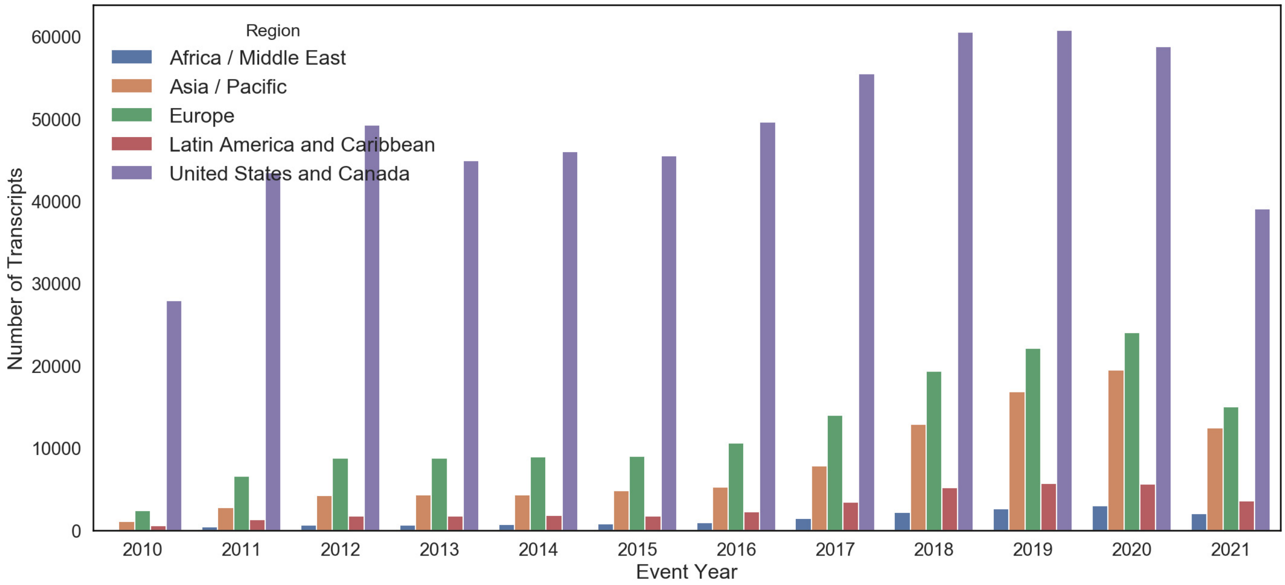


Exhibit 3 shows the number of transcripts for the calls in Exhibit 2. Our data provider retains multiple versions of each transcript depending on creation times. The earliest versions are created immediately following earnings calls but may contain typographical and transcription errors. Edited versions are typically available within one to two days after the call and are generally of higher quality. The data provider also randomly samples and audits some of the copies—while these versions tend to be the most accurate, they are generally available days or weeks after the actual call. To reinforce this point, since we cut our data to include available transcripts for calls that occurred up to and including December 31, 2021, there are less transcripts in 2021 when

Exhibit 3. Total number of earnings call transcripts by region by year based on event time.



compared to the prior years but we expect the data provider to add additional transcripts during 2022 for events prior to 2022.

Company earnings are typically reported quarterly so we analyze the availability of transcripts by month. Exhibit 4 shows the percentage of transcripts available each month based on their availability rather than the timing of the associated earnings calls. Across the regions, most of the transcripts are delivered in February, May, August, and November, consistent with fiscal calendars and regulatory requirements. While the fiscal years of some companies do not correspond with the actual calendar years, many companies still follow the normal calendar and as a result are required to file their quarterly reports 40–45 days after quarter-end, and their annual reports 60–90 days after year-end depending on the size of the company.

Generally, calls last about one hour with some variation. Exhibit 5 shows the length of the calls by region. Although there are differences

Exhibit 4. Availability of transcripts by month by region based on transcript creation time.

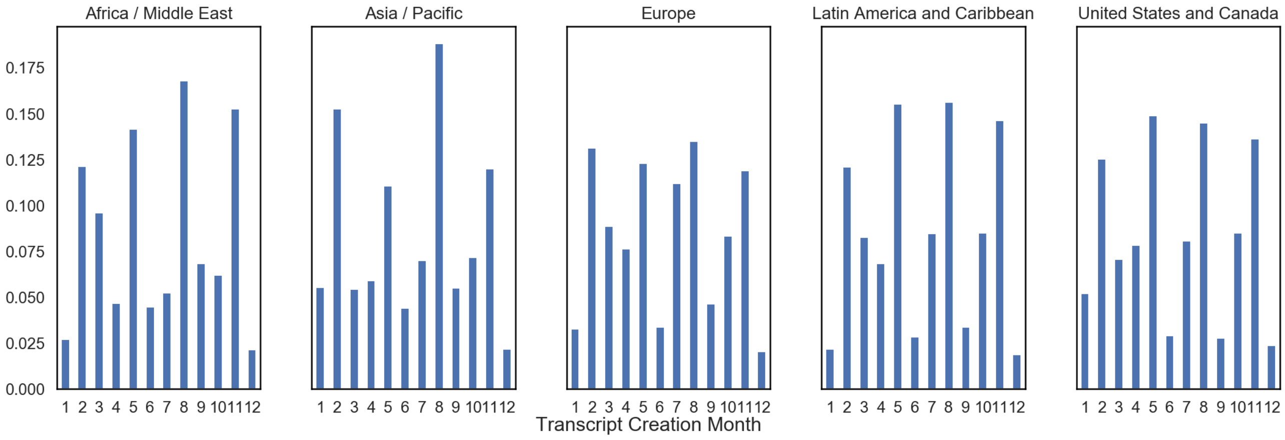
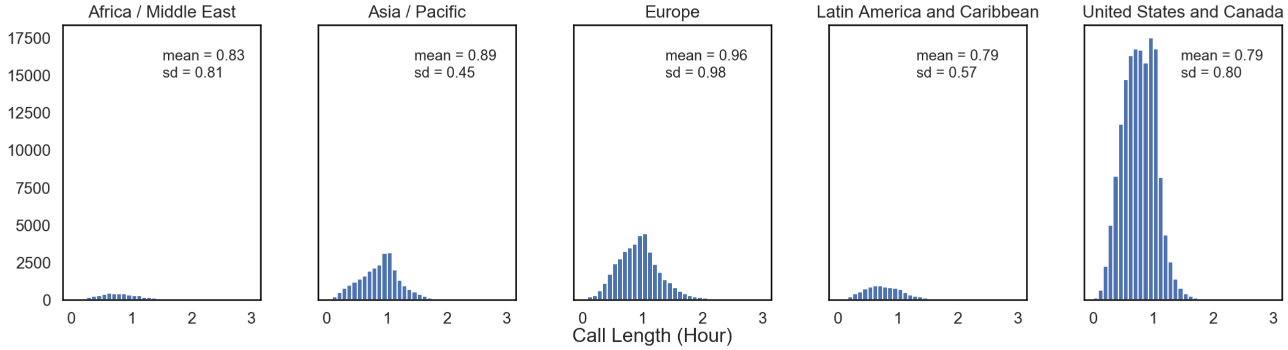


Exhibit 5. Length of calls by region.



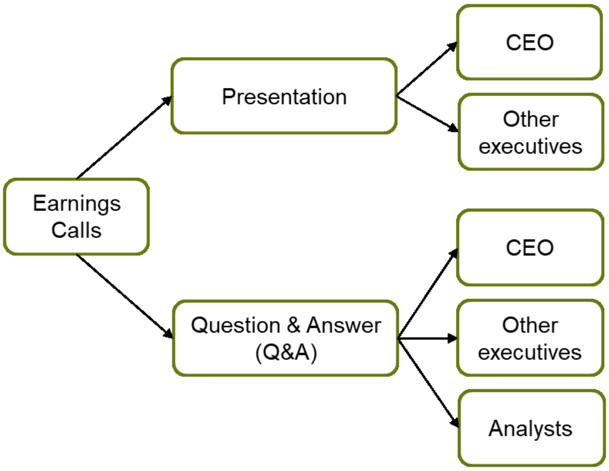
amongst the distributions of the call lengths, these variations are not significant.

3 Feature Generation

While human analysts may be able to study individual calls with enormous depth and synthesize information from financial results as well as the sentiment, machines are typically more effective at analyzing a large set of companies quickly. Machines can be taught to apply a set of rules to extract insights from earnings call transcripts. For example, analyzing the sentiment from company executives or analysts may provide useful information about a company's prospects. While company executives will generally try to present their financial results in the best light, we can leverage text-mining techniques to discern changes and trends. In addition, analysts may be seen as more objective, and their tone and sentiment may convey information about the future prospects of a company.

With these motivations in mind, we generate a wide variety of features to assess the potential of text-mining techniques to predict future stock returns. Our intent is not to optimize the features to generate a profitable trading strategy. Rather, we hope to demonstrate that there is ample opportunity for investors to apply NLP techniques in earnings call transcripts to differentiate between outperforming and underperforming stocks.

Exhibit 6. Earnings call sections.



We show a high-level overview of a typical earnings call in Exhibit 6. There is usually a presentation section and a question-and-answer (Q&A) section. Company executives are the sole participants in the first section while both corporate executives and analysts from the investment community interact during the Q&A section. The presentation section normally contains prepared remarks while the Q&A section allows for spontaneous discussion. We split the company executives into two groups, CEOs and Other Executives, to study potential differences between the two cohorts. In total, we split each earnings call into five individual sections, and we generate features within each of the sections as well as in combinations of these sections.

We calculate three broad categories of NLP features, namely, document attributes, readability scores, and sentiment scores.

3.1 Document attributes

Document attributes refer to features derived from the characteristics of the call. Examples include the number of words, sentences, questions, and analyst participants in a call. Given the importance of the CEO in a company, we measure various CEO characteristics such as the prevalence of adverbs in the CEO section, the proportion of plural pronouns within all pronouns used by the CEO, and the degree of CEO involvement using the percentage of the CEO's words within the sections containing executive comments. As an example, we show the distribution of the number of words in each of the five sections across various speakers in Exhibit 7. The lengths of CEO presentations and Other Executives presentations are similar. The length of CEO answers is longer on average, compared to the length of Other Executives answers. Note that since analysts are simply asking questions, they tend to use the least number of words.

3.2 Readability scores

Readability scores utilize a variety of methods to assess the difficulty of the text and document. These metrics tend to focus on two areas: the use of difficult words and the length of sentences. Easily understood messages (texts with low readability scores) may be quickly incorporated into market prices and will therefore have little impact on potential mispricing. On the other hand, complex messages (texts with high readability scores) may be used by company executives to obfuscate bad news or subpar results. Exhibit 8 shows the feature values of a sample readability score. On average, we found that the

Exhibit 7. Histogram of word counts across transcripts.

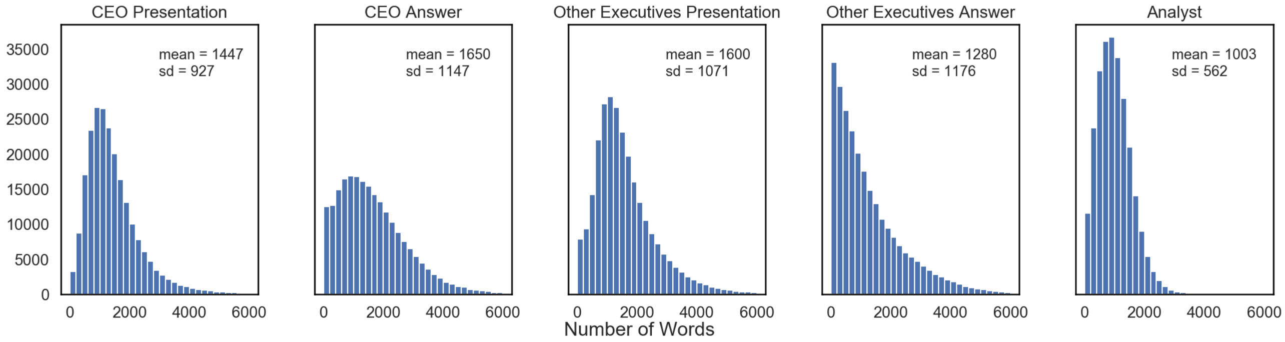
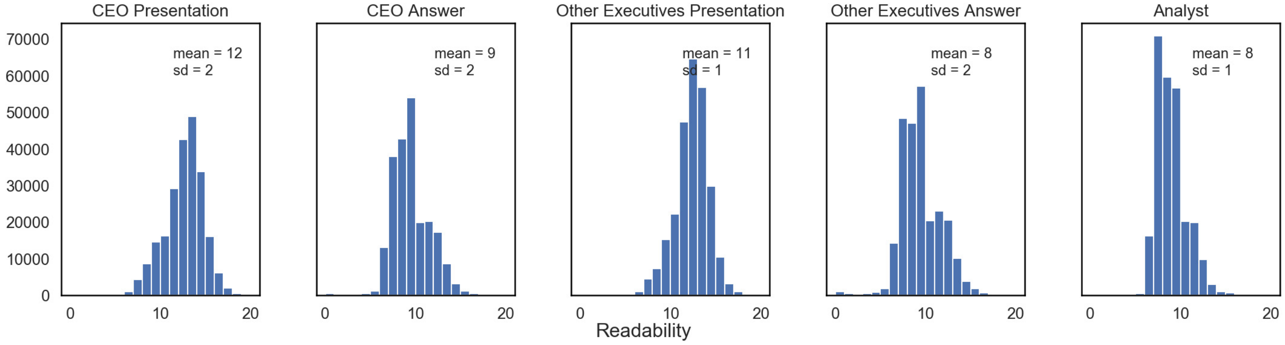


Exhibit 8. Histogram of one readability score.



presentation sections are the most difficult sections to understand. These sections are usually prepared remarks and may contain carefully constructed sentences. On the other hand, comments made in the Q&A section tend to be more informal and more direct.

3.3 Sentiment scores

Finally, we analyze different formulations of sentiment derived from the underlying text. The most basic method to assess sentiment is to count the number of positive and negative words based on a specific dictionary. As discussed previously, we calculate sentiment features using three dictionaries—Harvard IV-4, VADER, and Loughran McDonald (LM). Sentiment can be assessed using different methods. For example, we can calculate the net sentiment (positive score minus negative score) or the polarity (normalized net sentiment) of all the words in a document. Exhibit 9 shows the net sentiment of the CEO Presentation and the CEO Answer sections calculated using LM and VADER. From the charts, we note that there are more variations in the scores derived by the LM dictionary, consistent with the prior that this dictionary captures the financial usage of words and thus can differentiate sentiment more effectively.

Assessing sentiment with the BoW approach has limitations. For example, while "resolve" will generally be scored as a positive term under most dictionaries, the context around the word "resolve" is just as important. The phrases "did not resolve" and "need to resolve" are generally scored as negative by humans but BoW may not understand the nuances of the expressions. As a result, context-driven approaches to understanding language can potentially enhance the accuracy of sentiment scores.

Researchers have found that context-driven language models like BERT can perform well in common language tasks (Devlin *et al.*, 2019). Araci (2019) showed that the BERT base model can be leveraged to derive sentiment for sentences in financial documents. In our research, we modify Araci's approach in two important ways. We replace the BERT base model with the distilBERT model (Sanh *et al.*, 2019) to speed up model computations. This modification halved the processing time needed to manipulate and score the historical transcripts. We also fine-tuned the distilBERT model to improve the performance of the down-stream sentiment classification task with three datasets. The first dataset contains a subset of the labeled sentences from Financial PhraseBank (Malo *et al.*, 2014), the main data used in Araci's study. However, we only use the 2,262 sentences that were labeled consistently by the 16 annotators with backgrounds in finance and business because we believe that the resulting sentiments of these sentences are likely meaningful. The second dataset includes 1,100

Exhibit 9. Histogram of feature values for CEO sentiment using LM and VADER.

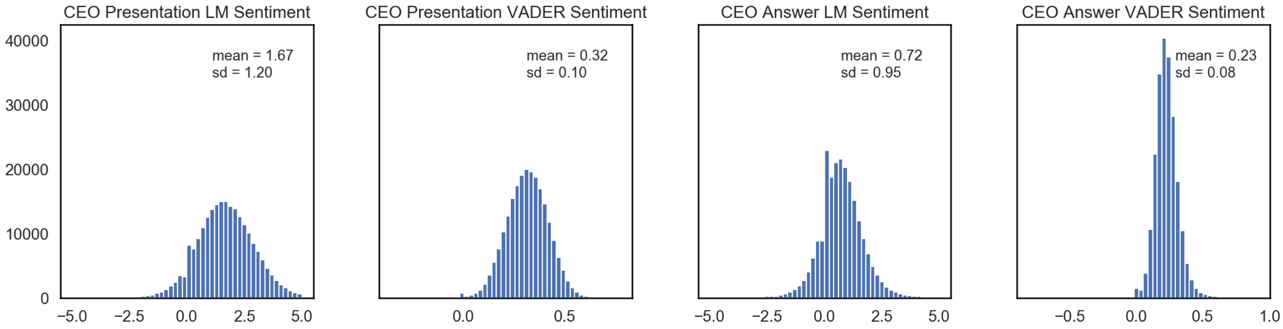
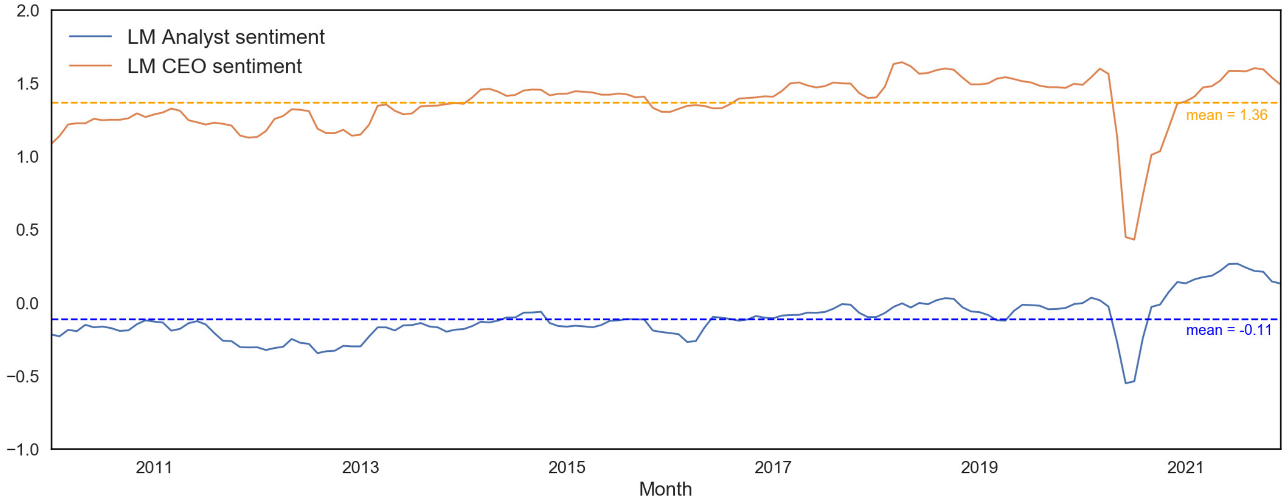


Exhibit 10. Monthly mean of the LM CEO Presentation and Analyst net sentiment for USLC stocks.



sentences from three transcripts prior to 2010 that we manually labeled to specifically improve the model's application in earnings call transcripts. The third dataset includes 120 greeting sentences that we labeled as "neutral" to mitigate their impacts on the sentiment scores. Examples include "Good day, everyone." and "I am very happy to take questions." With our fine-tuning, the resulting model is not only faster than Araci's BERT model, but also performs better in deriving sentiment from earnings calls transcripts. At the risk of creating confusion, we call our resulting model "finBERT" because we view this name as a generic term that refers to the application of the BERT model in finance. Note however, that our finBERT is significantly different from Araci's finBERT, as described above.

We apply our finBERT model in the earnings call transcripts to predict the probabilities of each sentence to be "Positive", "Negative", or "Neutral". We calculate the net sentiment of a sentence as its probability of being "Positive" minus its probability of being "Negative". We then construct finBERT net sentiment scores for each transcript by calculating the mean net sentiment across all sentences in each section.

We now provide some examples of sentiment features to show their trends and distributions over time. Exhibit 10 shows the mean of the LM net sentiment scores of the CEO presentation and Analyst question sections for all the stocks in the US Large Cap (USLC) universe. Based on the plot, there is a steady upward trend in the CEO net sentiment mean, suggesting that on average, CEOs at large US companies have been speaking in more positive tones over the recent period. This is consistent with other research findings which suggest that CEOs are changing their behavior in earnings calls because they are reacting to the recent trend of machines reading and analyzing their words. The study by Cao *et al.* (2021) is one example of this line of research. Note there was a large dip in CEO sentiment in June 2020 and July 2020 due to the impact of COVID-19 and its global economic aftershocks on company financials. This exhibit gives us confidence that the feature is capturing the language and sentiment used by the CEOs over the study period.

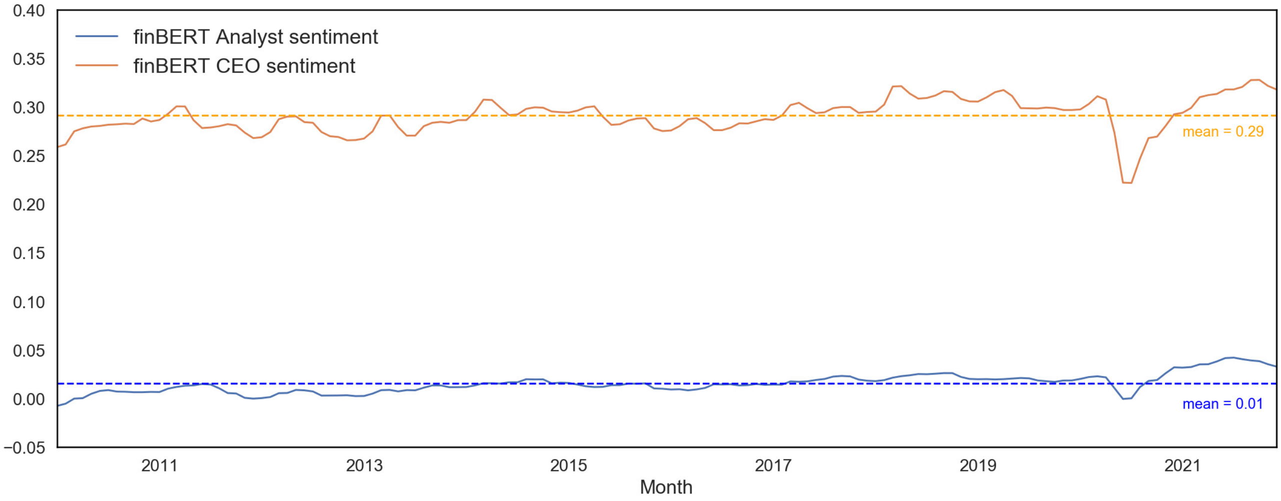
Compared to the average CEO net sentiment, it is interesting to note that the average Analyst net sentiment is much lower. Over our study period, the mean CEO net sentiment is 1.36, whereas the mean Analyst net sentiment is closer to zero at −0.11. In addition, the average Analyst net sentiment did not rise significantly over the study period, suggesting that analysts have not changed their language or style in the face of automation and text-mining applications. For us, going forward, this may suggest analyst-based features may be more objective and more enduring in their predictive capabilities when compared to CEO-based features.

Exhibit 11 show the monthly mean of the finBERT CEO and Analyst net sentiment, respectively, for USLC stocks. Similar to the LM net sentiment exhibits, the average finBERT CEO net sentiment (0.29) is much higher than the average finBERT Analyst net sentiment (0.01). In addition, CEO net sentiment increased more than Analyst net sentiment before COVID-19. Finally, we note that there is more variation in the CEO net sentiment scores, suggesting that Analysts tend to be more consistent in their comments.

In total, we create 280 features for the five individual sections (CEO presentation, CEO answer, Other Executives presentation, Other Executives answer, and Analyst questions) and three combined sections (CEO, Other Executives, and all Executives) for each transcript. In aggregate there are 40 document attributes, 64 readability scores, and 176 sentiment scores. These features can be found in the Appendix. We view these features as the basic foundations for NLP research—we expect asset managers to incorporate their proprietary insights to customize the features used in stock selection and portfolio construction.

We build a robust and scalable pipeline to facilitate and automate data ingestion, text cleansing, and feature engineering. We store all the transcripts in a MongoDB database and parse each transcript to extract the text from the five individual sections and retain the speaker tags. We then apply various text cleansing techniques to prepare the corpus for feature engineering. These techniques include the removal of special characters and stop words, tokenizing the text into words and sentences, and lemmatization. After cleansing the text, we apply our feature engineering methods across all the transcripts. Our infrastructure leverages modern technologies like Airflow, Azure, and Spark.

Exhibit 11. Monthly mean of the finBERT CEO Presentation and Analyst net sentiment for USLC stocks.



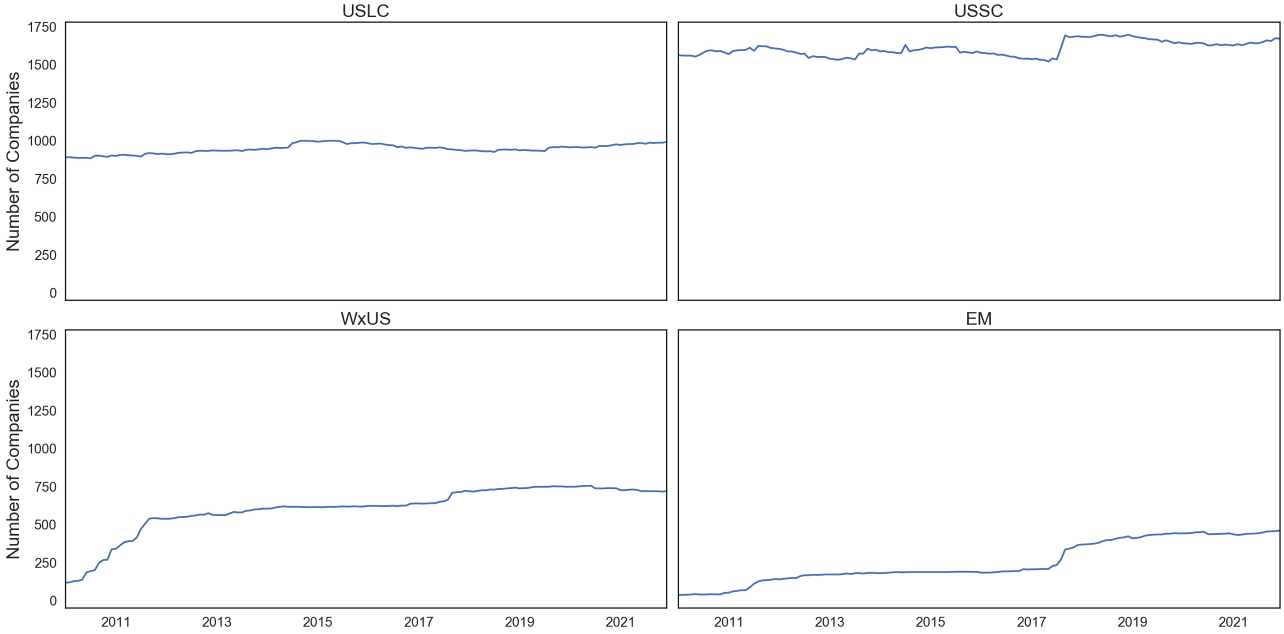
4 Backtest Methodology

We now test our generated features for stock selection and assess the ability of our features to differentiate between outperforming and underperforming stocks. For our backtests, we rebalance the portfolio every month using the latest version of each transcript for all the stocks in our universe. Note that we are not executing our strategy immediately after earnings releases and allow for days or weeks to pass before utilizing the generated signals. Thus, we are simulating a realistic implementation for the investment community.

We use a common long/short backtesting approach in the industry to test the predictability of the features. At the beginning of each month, we form portfolios based on the available features as of that date and track the difference in subsequent returns between the top and bottom quintiles over the following month. Specifically, for each feature, we extract the latest feature value for all the stocks in the universe and then rank all the stocks by the feature value. We then split the stocks into quintiles, with Q1 and Q5 representing the lowest and highest 20% of the stocks, respectively, based on the feature values. We rebalance the quintiles monthly and track the performance of each quintile over the following month. Note that at quintile formation each month, we only use data and features that were available at that point in time. We track the performance of each quintile over the full study period. In assessing a feature's effectiveness, we expect the top and bottom quintiles to exhibit persistently different return patterns over time.

We test each of the 280 features from 2010 to 2021 in four stock universes: US Large Cap (USLC), US Small Cap (USSC), World ex-US (WxUS), and Emerging Markets (EM). We derive these universes from the broad indexes used in the respective markets. For example, for our USLC universe, we map all the companies with earnings call transcripts to the point-in-time stock constituents of the Russell 1000 index. The resulting USLC dataset on average covers about 95% of the Russell 1000 by market capitalization over our study period but the coverage is lower for the other

Exhibit 12. Number of companies by universe.



universes. The size of our dataset by universe is shown in Exhibit 12. As expected, the coverage is much higher within the US but even outside the US, we believe that there are a sufficient number of companies to conduct our research.

5 Backtest Results

Given the breadth of features, we summarize our results from a high-level perspective rather than show the results of each individual feature. Our goal is not to determine the most efficacious features in our study period but to present an overview of the effectiveness of NLP-based signals for investors. We use the information ratio (IR) of a strategy as a representative risk-adjusted performance metric. The IR, defined as the annualized excess return divided by the annualized volatility of those returns, is a common metric used by the industry to assess factor efficacy. Equation (1) shows the calculation of IR.

$$IR = \frac{R_p - R_b}{\sigma_{(R_p - R_b)}}, \tag{1}$$

where R_p is the return of portfolio, R_b is the return of benchmark, and $\sigma_{(R_p - R_b)}$ is the tracking error of the portfolio returns versus the benchmark returns. For example, if the top quintile of a feature has an annualized return of 10% while the broad market returned 8%, the excess return of the strategy is 2%. Furthermore, if the annualized volatility of the relative returns is 5%, then the IR is 0.4 (2%/5%). Note that in this example, the absolute return of the top quintile reflects the returns from the broad market as well as the underlying feature.

In our study, we assess the efficacy of each feature by simulating a strategy where we buy the top 20% and sell the bottom 20% of stocks as ranked by the feature. This long/short strategy removes the impact of broad market movements and isolates the performance of the underlying feature. Since there is no broad market impact, we calculate the IR of this strategy by comparing it with zero. While different practitioners have different thresholds for determining significance, information ratios above 0.5 are generally considered worthwhile for further research. For our study, we use a slightly higher threshold based on the following calculations. With 11 years of history in our dataset, we calculate that IRs above 0.58 are statistically significant if we assume that IRs follow the t-distribution. Specifically, the minimum one-tailed t-statistic that is significantly different from 0 at the 95% confidence level with 10 degrees of freedom (11 years of data) is 1.82. Thus, if $IR > 0.58$, Equation (2) is satisfied,

$$\frac{R_p - R_b}{\frac{\sigma_{(R_p - R_b)}}{\sqrt{11-1}}} = IR \times \sqrt{10} > 1.82. \tag{2}$$

We evaluate all features in four universes: USLC, USSC, WxUS, and EM. Other text-mining studies have generally focused on US large-cap universes, but our study extends the research into other universes to test for robustness across a wide variety of stocks.

We show a histogram of the full period IRs for all the features for each universe in Exhibit 13. Note that the exhibit shows the absolute value of the IRs since negative IRs from a Q1–Q5 analysis can simply be transformed to positive IRs by using Q5–Q1. This transformation accounts for the fact that the most attractive stocks may reside in Q5 rather than Q1, depending on the feature. From the exhibit, approximately 10–20% of the IRs are greater than 0.58 for the USLC, WxUS and EM universes and around 30% of the IRs are above this threshold for the USSC universe. These high-level results suggest that our generated features may be useful for active managers across a broad set of strategies.

Next, we compare the performance of the three broad categories of features. Exhibit 14 shows the box plots of the absolute IRs for features within

Exhibit 13. Distribution of absolute IRs for all features.

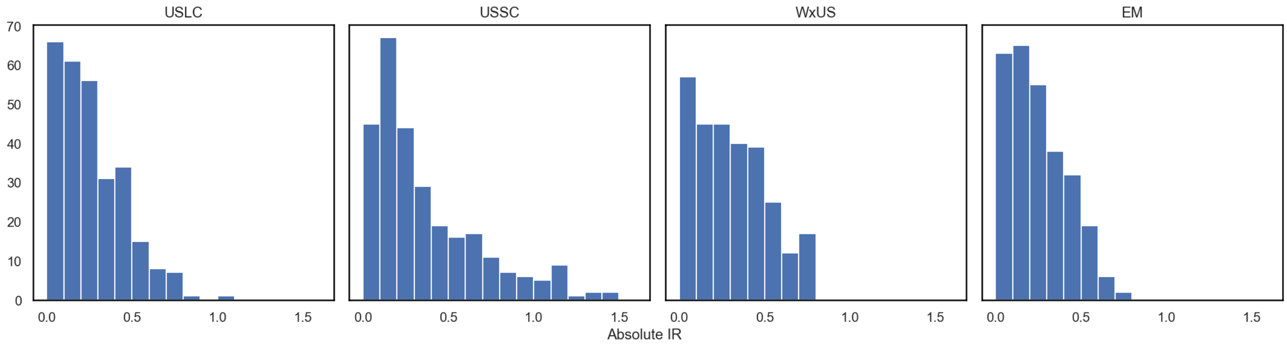
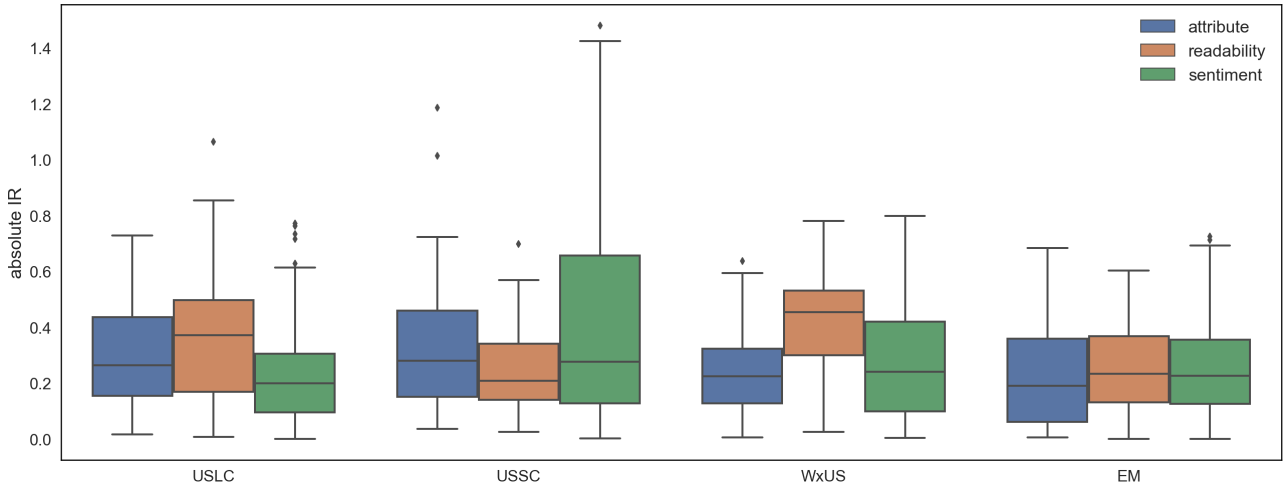


Exhibit 14. Absolute IRs of the different categories of NLP features across universes.



document attributes, readability, and sentiment across the universes. Consistent with Exhibit 13, the sentiment features for the USSC universe have consistently higher IRs than the other universes. Note also that the readability features may hold the most promise for USLC and WxUS since around 17% (23%) of the readability features for USLC (WxUS) have absolute IRs above 0.58.

We now delve further into the sentiment-based features because of their potential to capture linguistic characteristics of the words and sentences used by the speakers. We evaluate their efficacy along three dimensions. First, we compare the performance of the individual sections of the CEOs and Other Executives with a combination of these speakers. While CEOs generally speak during conference calls, Other Executives may play more prominent roles for some companies. For example, the COO and CFO are the most prominent speakers for companies like Dell Technologies during our study period. To capture "executive sentiment" more broadly throughout a call, we combine the sentiment features from all the executives (i.e., CEOs and Other Executives) across both the presentation and Q&A sections. We want to analyze the benefits of combining speakers and sections to determine if the resulting features are more robust.

We then compare the performance of the sentiment features from executives and analysts. We believe that commentary from executives and analyst questions carry different and potentially contrasting insights, especially since our findings earlier in the paper suggest that company executives are changing the way they communicate in earnings calls.

Finally, we compare the performance of the LM dictionary sentiment features and the finBERT sentiment features. There is evidence that the former works well in financial documents (Loughran and McDonald, 2011), so it serves as a good baseline to understand the performance of the latter.

Exhibit 15 shows the IRs of CEO net sentiment in the Presentation and Q&A sections as well as the Other Executives net sentiment in the same two sections, based on the LM dictionary. The performance patterns of the CEO net sentiment in both the two underlying sections are relatively consistent across the universes while the results of

Exhibit 15. IR of Q1–Q5 for LM net sentiment from CEO Presentation, CEO Answer, Other Executives Presentation, and Other Executives Answer across universes.

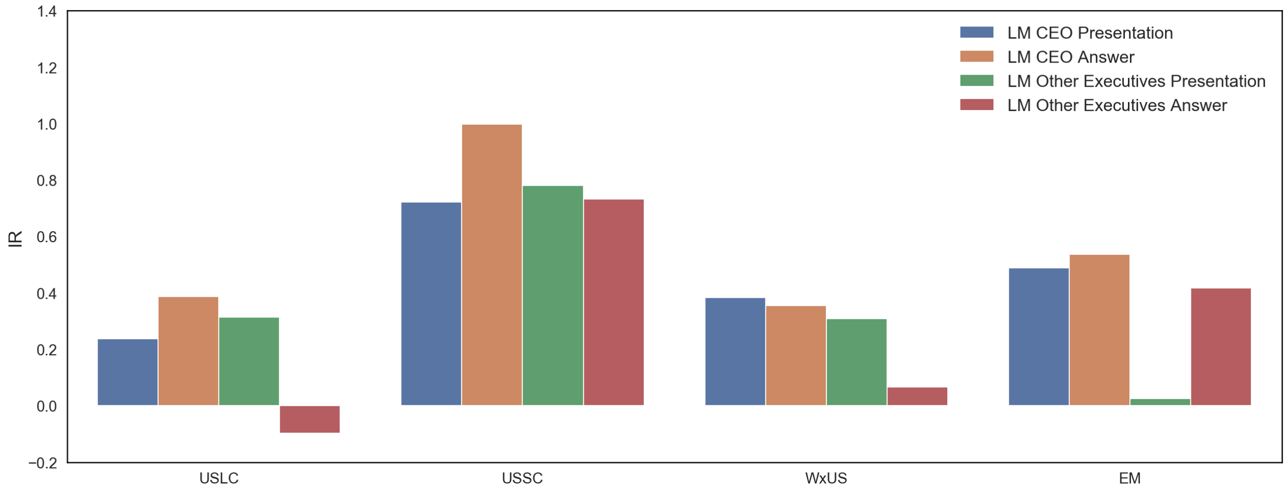
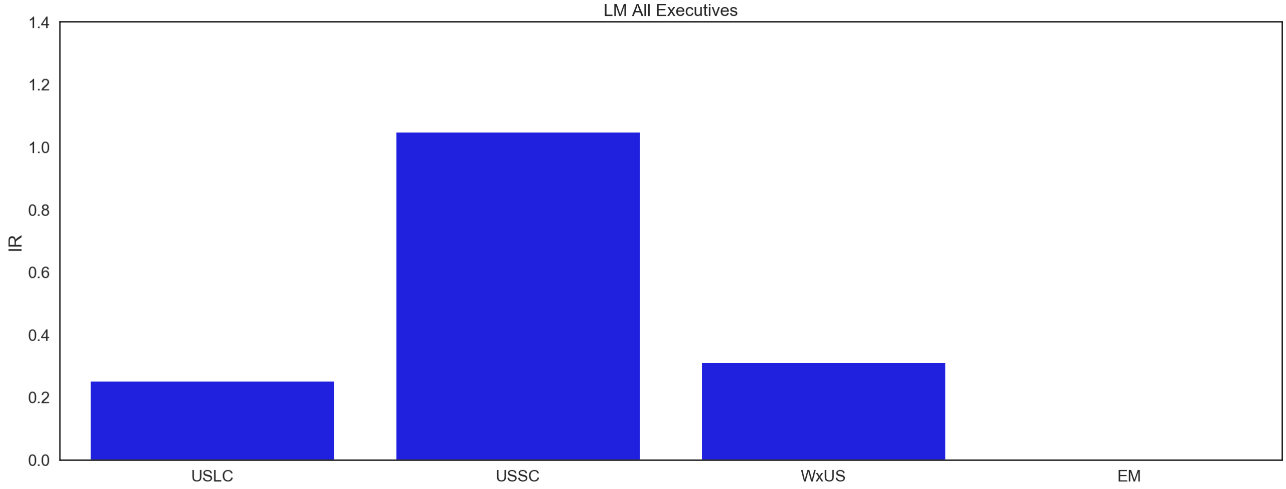


Exhibit 16. IR of Q1–Q5 for LM net sentiment from All Executives across universes.



the same features for Other Executives are more volatile. Further analysis is required to understand the differences but that is beyond the scope of this paper.

Exhibit 16 shows the IRs of All Executives net sentiment for the whole document over the same period as Exhibit 15. For the USSC universe, note that the IRs of the combined All Executives feature are as strong as the individual CEO and Other Executives signals. Interestingly, for the other universes, the combined All Executives signal is not as strong as the average of the underlying signals, implying that it may be more advantageous for investors to leverage the features from the individual speakers.

Exhibit 17. Dollar growth based on LM net sentiment from CEO Presentation, CEO Answer, Other Executives Presentation, Other Executives Answer, and All Executives across universes.

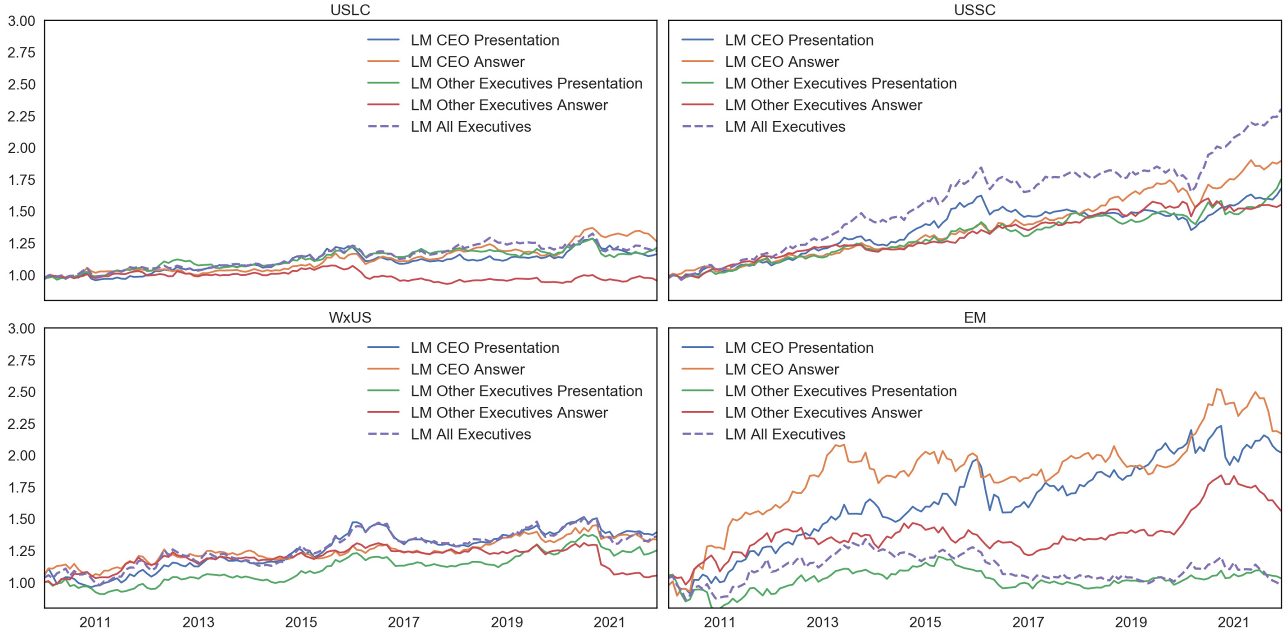


Exhibit 18. IR of Q1–Q5 for LM net sentiment from Analyst across universes.

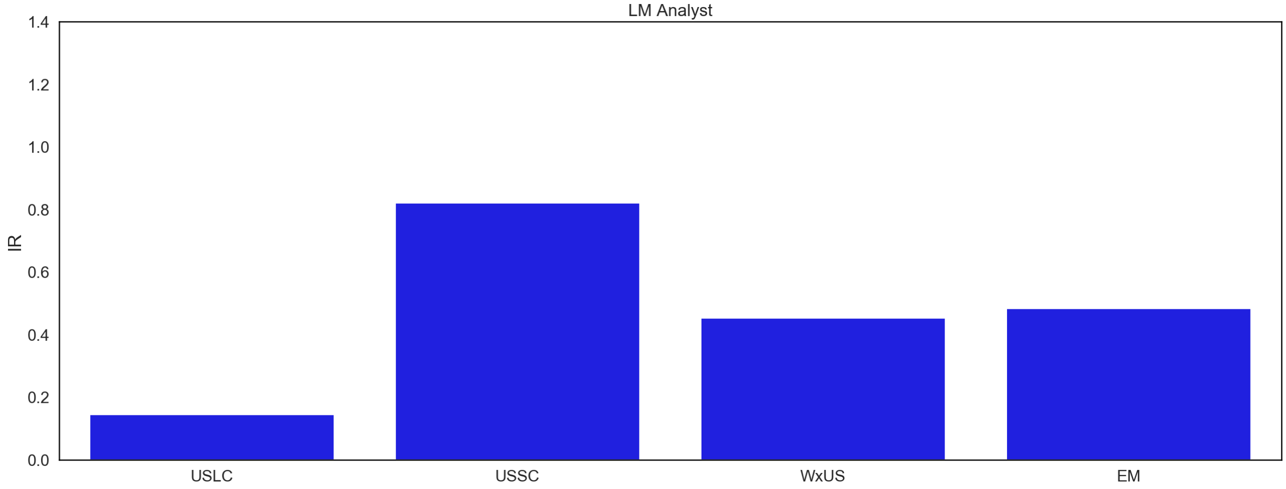


Exhibit 17 shows the dollar growth based on LM net sentiment from the individual sections and the combined sections. Consistent with the results from Exhibits 15 and 16, the net sentiment from the All Executives section outperforms all the individual sections for the USSC universe. For the USLC and WxUS universes, the combined feature is comparable to the best performing individual features. On the other hand, the combined sentiment signal is weaker in EM. One explanation for this poor performance may be our dataset's relatively low coverage of this universe.

Exhibit 18 shows the IRs of Q1–Q5 for LM net sentiment from analyst questions. Compared to the same feature generated in All Executives in

Exhibit 19. Dollar growth based on LM net sentiment from Analyst across universes.

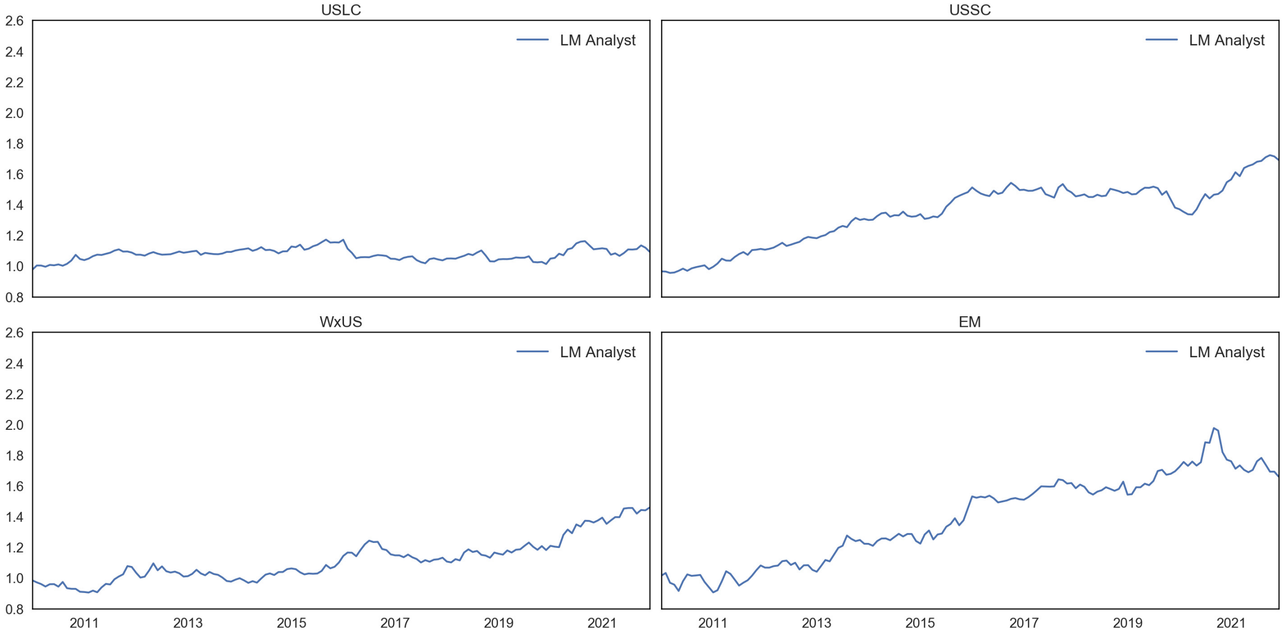
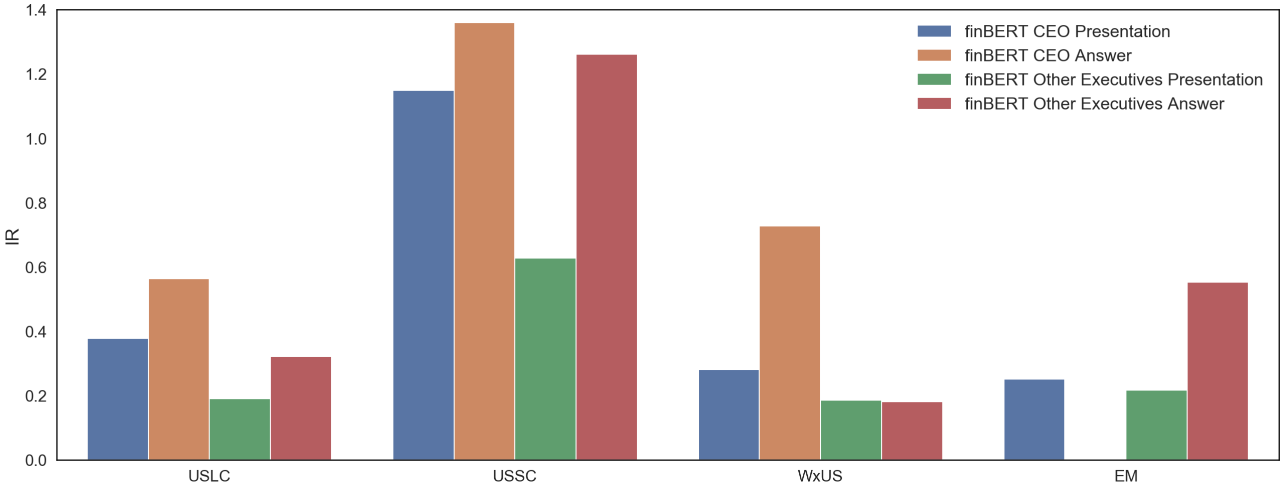


Exhibit 20. IR of Q1–Q5 for finBERT net sentiment from CEO Presentation, CEO Answer, Other executives Presentation, and Other Executives Answer across universes.



the previous exhibits, these results are similar in USLC, USSC, and WxUS, and much stronger in EM. The EM results suggest that the analyst sentiment provides a promising signal for investors in spite of the lackluster results using the executive sections. The corresponding dollar growth charts are shown in Exhibit 19.

We now examine the performance of the finBERT net sentiment in the various speakers and sections of the earnings call transcripts.

Exhibit 20 shows the IRs of Q1–Q5 for finBERT net sentiment from individual sections and Exhibit 21 shows the results for All Executives.

Exhibit 21. IR of Q1–Q5 for finBERT net sentiment from All Executives across universes.

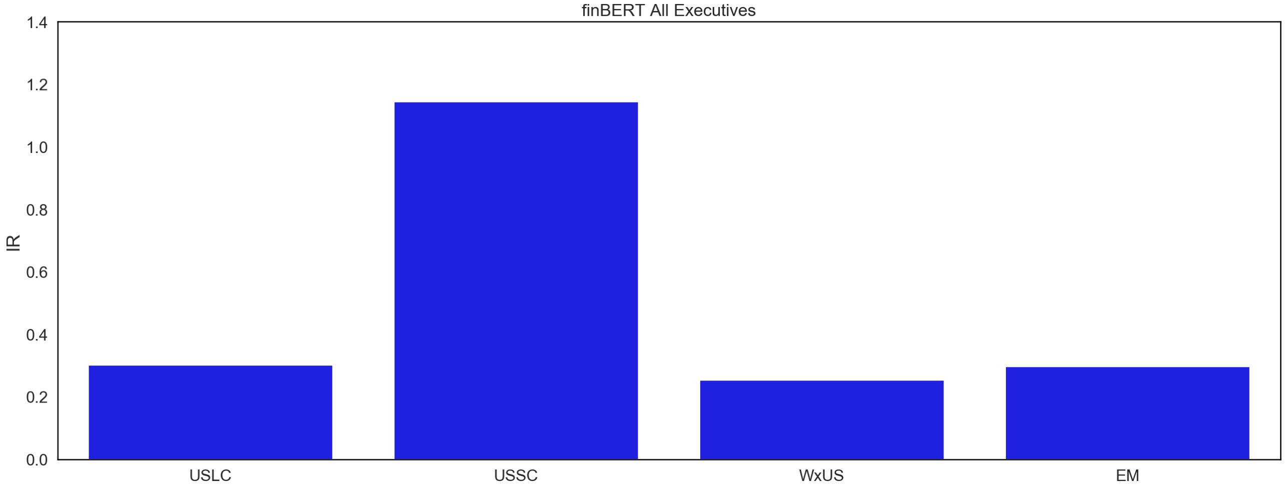
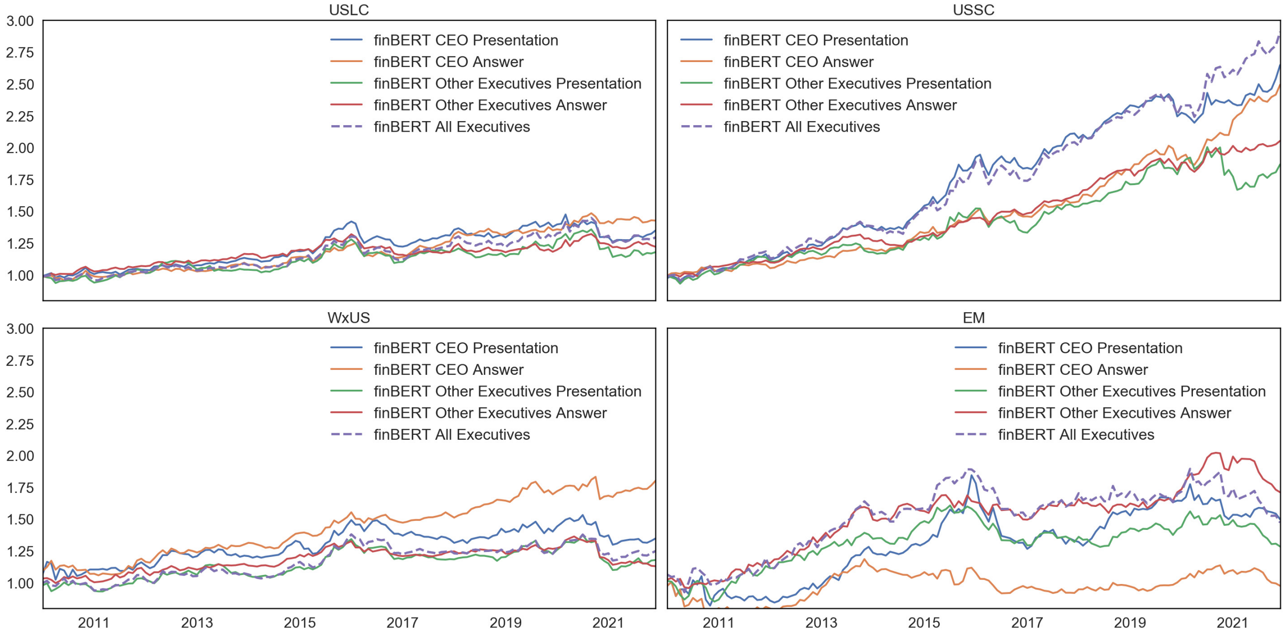


Exhibit 22. Dollar growth based on finBERT net sentiment for CEO Presentation, CEO Answer, Other Executives Presentation, Other Executives Answer, and All Executives across universes.



Across the universes, the performance of the finBERT net sentiment from combining the sections is similar to the average of the results in the individual sections, suggesting that the comprehensive finBERT features capture the collective sentiment of the various speakers throughout the calls. The dollar growth charts in Exhibit 22 lead to the same conclusion.

Exhibit 23 shows the IR of the finBERT Analyst net sentiment and Exhibit 24 shows the corresponding dollar growth charts. Compared to the exhibits for All Executives sentiment, these Analyst exhibits suggest the analyst section may offer more compelling insights into the USLC and WxUS universes and comparable insights into the other two universes. Consistent with our previous comments, we find that Analysts tend to be more consistent in their comments and sentiment, and as a result, features generated in Analysts may provide more discerning signals.

Exhibit 23. IR of Q1–Q5 for finBERT net sentiment from Analyst across universes.

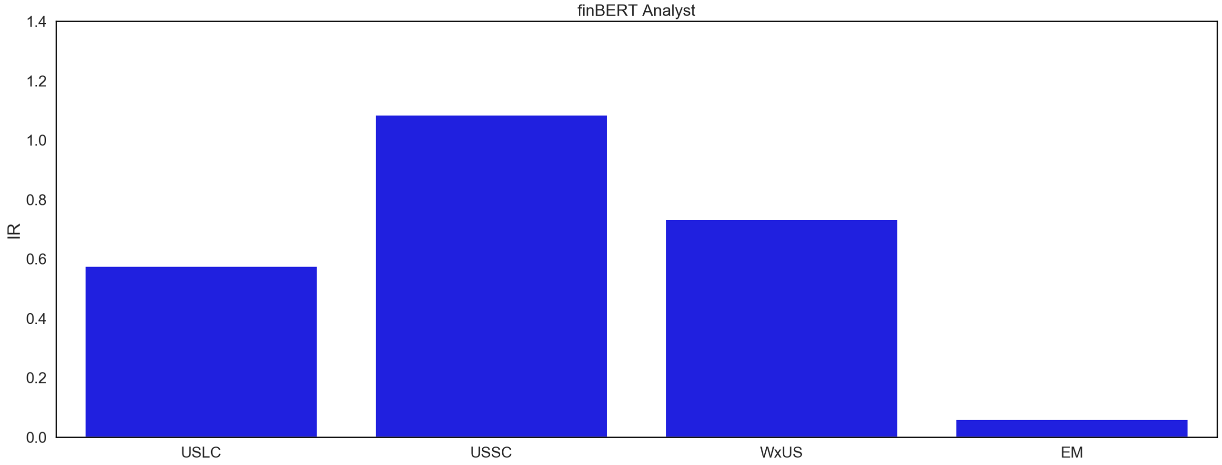
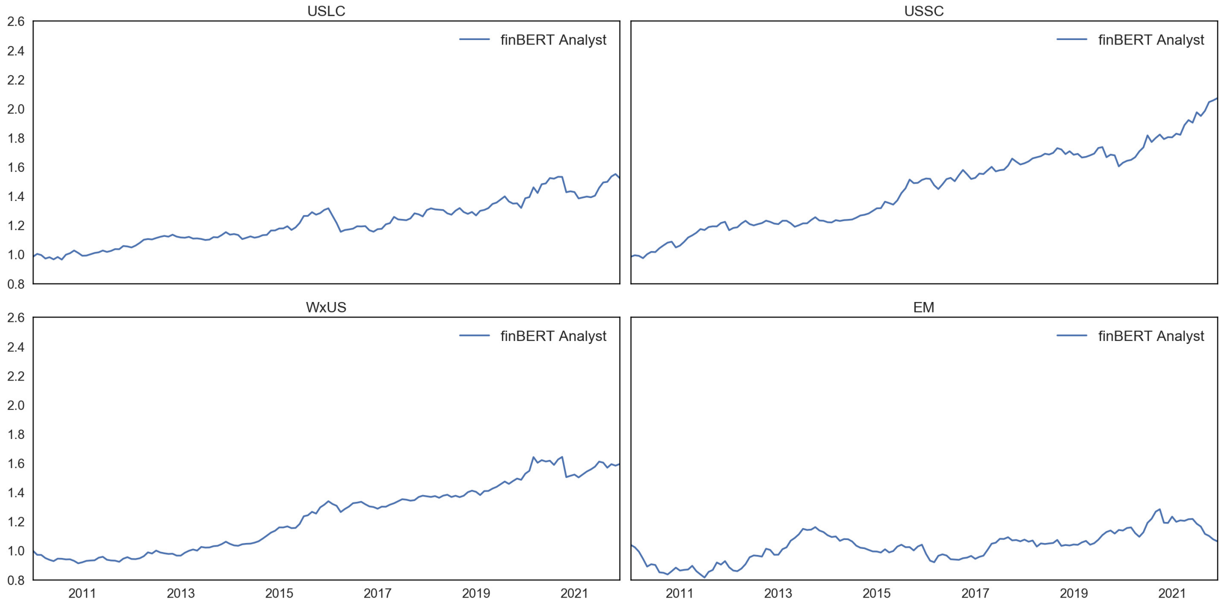


Exhibit 24. Dollar growth based on finBERT net sentiment from analyst questions across universes.



We now compare the performance of the net sentiment scores between LM and finBERT shown in the previous exhibits. Our prior is that context-driven approaches like finBERT should outperform BoW approaches like LM. In general, our results are consistent with that prior. For both US large-cap and small-cap universes, the performance of the finBERT derived features is stronger than the results from the LM-derived features. Within the World ex-US universe, the results are similar using the two methods for executive sentiment, but the performance of analyst sentiment using finBERT is stronger than LM. The one area where our results are not intuitive is the emerging markets universe—here, the LM-derived features performed better historically. As discussed earlier, our coverage in the EM markets is limited so that it may contribute to the different results. In addition, there may be language-related reasons for this underperformance. For many of the executives leading the emerging market companies, English is not their native language so it may be more difficult for them to fully express their views. In addition, models like finBERT may not be able to pick up the nuances in the text because these executives may be using words, phrases, and sentences differently from native English speakers. These hypotheses need more investigation.

6 Conclusion

We generate a representative set of features in earnings call transcripts using NLP techniques and show that there are opportunities for text mining in asset management. Specifically, we create document attribute, readability, and sentiment features from the transcripts and find that the last two types of features can help investors differentiate between outperforming and underperforming stocks across different universes. We find that more than 20% of the features are promising across the four universes, suggesting that sorting stocks by our NLP signals can be rewarding for a wide range of investment strategies. Our results demonstrating the efficacy of NLP-based features amongst small-cap and non-US companies are novel to the existing literature. We believe that asset managers can create even more powerful features by integrating their fundamental priors into signal construction. While our findings are particularly strong in the US small-cap universe, there are numerous opportunities to outperform in the other markets using NLP features. We also find that the sentiment features are robust across different speakers and different sections of the earnings calls. The evidence on combining features across speakers and sections is not conclusive but suggests that investors should take special care in their signal construction.

In addition, our research suggests that utilizing analyst sentiment may be more rewarding than executive sentiment. We show evidence that executives are adapting to text-mining algorithms while analysts have been more consistent in their remarks and semantic usage over time.

Finally, we find that context-driven models like finBERT are promising and can outperform dictionary-based approaches in extracting sentiment from text. We believe that deep learning models like BERT will have broader applications in finance, extending beyond sentiment analysis to text summarization and theme extraction.

Our research can be leveraged by active managers to sharpen their research insights and enhance their skill level. By systematically applying text mining to earnings call transcripts, investors can extract broader and deeper insights that may lead to stronger investment performance.

Appendix—List of NLP features

Category (count)	Feature	Description
Readability (64)	Automated readability Coleman Liau Dale chall Flesch ease Flesch kincaid Gunning fog Smog index Overall	• Readability scores are calculated using the Textstat package (https://github.com/shivam5992/textstat). • The eight readability scores are calculated in eight sections: CEO Presentation, CEO Answer, Other Executive Presentation, Other Executive Answer, Analyst, CEO, Other Executives, and All Executives. There are 64 in total.
Attribute (40)	Number of words Number of sentences Number to words ratio	• These three features are calculated in eight sections, 24 in total.
	Proportion of plural pronouns within all pronouns Proportion of adverbs Number of analysts Number of questions	• These two features are calculated in seven sections (not in the Analyst section), 14 in total.
Sentiment (176)	LM positive LM negative LM uncertainty LM litigious LM superfluous LM interesting LM modal weak LM modal moderate LM modal strong LM constraining LM complexity	• The LM scores are calculated based on the Loughran−McDonald Master Dictionary https://sraf.nd.edu/loughranmcdonald-master-dictionary/ in eight sections, 112 in total. • The HIV4 sentiment scores are calculated based on the Harvard IV-4 dictionary https://raw.githubusercontent.com/hanzhichao2000/pysentiment/master/pysentiment/static/HIV-4.csv in eight sections, 40 in total.
	LM net sentiment	• Net sentiment is calculated using $positive - negative$.
	LM polarity	• Polarity is calculated as $\frac{positive-negative}{postive+negative}$.
	LM subjectivity	• Subjectivity is calculated as $\frac{positive+negative}{postive+negative+neutral}$.
	HIV4 positive HIV4 negative HIV4 net sentiment HIV4 polarity HIV4 subjectivity	

(*Continued*)

Category (count)	Feature	Description
	VADER compound mean	• The VADER sentiment scores are calculated using the package https://github.com/cjhutto/vaderSentiment, 16 in total.
	VADER compound sd	
	finBERT sentiment	• The finBERT sentiment score is calculated in eight sections, 8 in total.

Acknowledgment

We thank Sachin Chavan and Vinayaka Holla for their enormous contributions in the data engineering aspects of the research.

References

Araci, D. (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," Available at https://arxiv.org/abs/1908.10063.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2021). "From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses," Available at SSRN: https://ssrn.com/abstract=3840538.

Chebonenko, T., Gu, L., and Muravyev, D. (2018). "Text Sentiment's Ability to Capture Information: Evidence from Earnings Calls," Available at SSRN: https://ssrn.com/abstract=2352524.

Das, S., Goggins, C., He, J., Karypis, G., Krishnamurthy, S., Mahajan, M., Prabhala, N., Slack, D., Dusen, R., Yue, S., Zha, S., and Zheng, S. (2021). "Context, Language Modeling, and Multimodal Data in Finance," *The Journal of Financial Data Science* **3**(3), 52–66.

Devlin, J., Chang, M., Lee, K., and Toutanova, K., (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Available at https://www.aclweb.org/anthology/N19-1423.pdf

Engelberg, J. (2008). "Costly Information Processing: Evidence from Earnings Announcements," *AFA 2009 San Francisco Meetings Paper*, Available at SSRN: https://ssrn.com/abstract=1107998.

Garcia, D., Hu, X., and Rohrer, M. (2020). "The Colour of Finance Words," Available at SSRN: https://ssrn.com/abstract=3630898.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). "Text as Data," *Journal of Economic Literature* **57**(3), 535–574.

Hutto, C. J. and Gilbert, E. (2015). "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, Available at http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.

Klevak, J., Livnat, J., and Suslav, K. (2019). "A Practical Approach to Advanced Text Mining in Finance," *The Journal of Financial Data Science Winter* **1**(1), 122–129.

Loughran, T. and McDonald, B. (2010). "When Is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance* **66**(1), 35–65.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and 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.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). "A Primer in BERTology: What We Know About How BERT Works," *Transactions of the Association for Computational Linguistics* **8**, 842–866, Available at https://doi.org/10.1162/tacl_a_00349.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," In 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing -NeurIPS, Available at https://arxiv.org/abs/1910.01108.

Tetlock, P. C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the stock Market," *Journal of Finance* **62**(3), 1139–1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention Is All You Need," *Advances in Neural Information Processing Systems* 5998−6008, Available at https://arxiv.org/abs/1706.03762.

Keywords: BERT, earnings call transcripts, natural language processing (NLP)