Predicting Stock Prices: A New Approach to ML-Driven Sentiment Analysis

Jake Lees

Bachelor of Business (Honours)

2023

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text. I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student: Jake Lees[1]

Date: 24th November 2023

[1] **Acknowledgements:** I would like to thank my supervisors – Kenny Phua and Vitali Alexeev for their feedback and guidance throughout the completion of this research project. A special thank you to my industry mentor David Walsh, for his assistance in providing a practical perspective on this research. I am also very grateful and appreciative to my good friend Daniel Berson for his support and valuable contributions throughout the project.

Table of Contents

List of Tables

List of Figures

Abstract

This study advances sentiment analysis by developing new industry-specific dictionaries through machine learning (ML) techniques. This differs from the previous methods in the literature that do not differentiate between industries when creating sentiment dictionaries. This study shows that the variation in language used across industries affects sentiment classification, previously overlooked by the one-size-fits-all dictionary. When predicting stock price movements, we find that ML-based sentiment analysis achieves higher returns on industries with greater volatility, and more accurately predicts stock price declines. This research introduces a new framework for sentiment analysis, providing research relevant to both practitioners and the academic literature.

Keywords: measuring sentiment, machine learning, earnings calls

JEL Classifications: D82, G14

1. Introduction

1.1. Textual Analysis

Textual analysis has become increasingly popular due to advancements in computer technology, learning algorithms, and its application in profitable trading strategies. Textual analysis involves Natural Language Processing (NLP), analysing unstructured text such as media articles, financial reports, and social media to extract predictive measures and insights. Several studies have utilised sentiment analysis in financial and accounting problems, such as predicting market volatility and movements (Tetlock et al., 2008), analysing financial constraints (Bodnaruk et al., 2015), and studying the impact of investor sentiment on predicting S&P 500 price movements (Sun et al., 2016). As a result, numerous models have been created to comprehend how investor sentiment may affect markets.

1.2. Lexicon-Based Sentiment Analysis

This paper sets out to analyse earnings call transcripts during exchanges between management and analysts (Q&A), with the aim of composing new, industry-specific dictionaries for sentiment analysis. Sentiment analysis use these dictionaries to assess the tone of language in company reports, like earnings calls. Traditionally, these dictionaries were compiled manually, with humans assigning sentiment values to words. More recently, advanced algorithms analyse large amounts of textual data to help colour these sentiment dictionaries. Sentiment analysis has proven effective in predicting stock market reactions; positive sentiments often correlate with rising stock prices, while negative sentiments can lead to declines.

The central hypothesis of this research is that using dictionaries crafted for specific industries will enhance sentiment analysis for companies within those sectors, as opposed to utilising generic, one-size-fits-all dictionaries. To illustrate, sentiment analysis for a company in the healthcare sector should employ a specialised healthcare dictionary, anticipated to outperform a universal dictionary that does not differentiate between company sectors. This approach is not merely a modification but a rethinking—shaping a pathway to refine sentiment analysis and build dictionaries with heightened specificity and applicability.

Building on the advancements in learning algorithms, this research explores the machine-learning (ML) application to detect sentiment in earnings calls and create a profitable

trading strategy. Notably, we draw inspiration from the work of Garcia et al. (2023), who significantly advanced our understanding of the predictive power of financial language. They pioneered the creation of a sentiment dictionary through a ML approach, using stock price reactions during earnings calls as a guiding light. This paper also takes an ML approach; however, we separate dictionaries into eleven sectors for more accurate sentiment analysis. We validate that the ML-based dictionaries yield positive trading strategies when tested on their ability to predict stock price returns around earnings calls.

This research identifies effective applications for sentiment analysis and provides insights for practitioners in the field. We find that sentiment analysis significantly benefits industries characterised by high volatility. Notably, the top four most volatile industries in our out-of-sample data outperformed the four least volatile industries by an annualised alpha of 4.37%. Additionally, our results indicate that sentiment analysis is more precise in predicting negative transcripts, thus favouring shorting strategies for traders. In practice, shorting based on out-of-sample data shows an annualised outperformance of 3.2% for short strategies over long. This suggests that negative sentiment in documents has a more pronounced impact on stock price declines.

The dictionaries we create highlight phrases unique to specific sectors, along with universal terms across all industries. For instance, in developing the Health Care industry dictionary, we identified positive bigrams like ‘research advancement,’ ‘flu vaccine,’ and ‘health economics’. These terms are unique to the Health Care industry and are not captured using the more generic dictionaries previously used in the literature. This new approach allows for a more accurate and specialised sentiment assessment, enhancing the precision of predicting stock price reactions for companies within specific industries.

As the field of textual analysis in finance continues to grow, it continuously integrates more sophisticated NLP and ML techniques to enhance sentiment analysis. Despite these advancements, using pre-specified sentiment dictionaries, such as the Harvard-IV, remains a core empirical method in finance studies. The Harvard-IV dictionary is based on the psychology literature, where humans label words (tokens) as positive or negative to create sentiment dictionaries. Tetlock (2007) applies the Harvard-IV dictionary to Wall Street Journal articles and finds that more pessimistic language generally points to a slight decrease in stock prices. These dictionaries were further refined by Loughran and McDonald (2011), who adjusted mislabelled tokens by utilising 10K financial statements. For instance, "liability" was initially

labelled as negative, yet its frequent use in balance sheet discussions does not inherently carry a negative sentiment. These studies suggest that lexicon-based sentiment analysis is a statistically significant predictor for stock returns, laying the foundation for future studies in this field.

However, these dictionaries have faced criticism for its inherent human bias and lack of predictive power compared to ML approaches. A more recent study demonstrates that the new ML dictionaries provided a more accurate and nuanced perspective of financial sentiment analysis and are superior in predicting out-of-sample stock price movements (Garcia et al., 2023). Our paper contributes to the debate by constructing new dictionaries using techniques from the NLP literature and comparing their performance versus previous dictionaries in out-of-sample tests. Our findings indicate a significant outperformance by the ML-based approach, achieving an annualised alpha of 8.52% over the human-based dictionaries, which returned 3.44%.

Our paper focuses on the Q&A sections of earnings calls and analysing the management responses (answers) to help create the dictionaries. This differs from Garcia et al. (2023), who developed dictionaries using whole 10-K statements and WSJ articles. The Q&A section is characterised by a high signal-to-noise ratio – a crucial factor for successful ML applications (Matsumoto et al., 2011). The tone of the Q&A portion of earnings calls is a significant predictor of abnormal returns. Their study suggests separating managerial and analyst contributions as an avenue for future research (Price et al., 2012). Building on this, we hypothesise that management's language has a more pronounced effect on stock prices, making it particularly valuable for developing our sentiment dictionaries. Therefore, we concentrate on management answers to refine the predictive quality of the ML approach and help colour the finance words in our dictionaries.

We use TF-IDF and Multinomial Naïve Bayes Classifier (MNB) as the methods to score the bigrams and help colour the dictionaries. Garcia et al. (2023) utilise the multinomial inverse regression model (MNIR) which is based on the framework used by Taddy (2013). However, this algorithm has been improved and extended, showing that other ML techniques are likely to produce similar or better results (Kelly et al., 2018). Bachhety et al., (2018) argue that MNB performs better than discriminative models such as SVM and decision trees for sentiment analysis tasks. MNB is used due to its strong text classification performance, driven by its ability to manage large-dimensional datasets.

Our algorithm's primary output is a collection of positive and negative bigrams, each with a corresponding score, which comprise the dictionaries. Following pre-processing and cleaning of transcript data, we constructed a universal dictionary by analysing 63,345 earnings call transcripts using our algorithm. We apply the same technique for industry-specific dictionaries but limit the data to transcripts from a single industry, allowing for more precise sentiment classification.

1.3. Aim of Research

Our research investigates whether industry-specific dictionaries outperform previously used general dictionaries in sentiment analysis. To validate our approach, we create a trading strategy using out-of-sample data to evaluate the effectiveness of these new dictionaries in predicting returns. By analysing unique jargon and connotations inherent to different industries, we aim to establish that these specialised dictionaries offer more nuanced insights into sector-specific language.

In developing industry-specific dictionaries, we identified unique bigrams that more accurately represent industry-specific terminology. For instance, in the Information Technology (IT) sector, we discovered that 36% of bigrams unique to this industry are not included in the universal dictionary. Analysis of these bigrams reveals that general dictionaries can overlook key terms necessary for assessing sentiment in the IT industry. For example, the top three positive unique bigrams identified in IT—'technology inflection,' 'early adopter,' and 'customer subscription'—are terms that can help for more accurately scoring documents for IT companies, yet they are absent in a universal dictionary.

Furthermore, our analysis reveals that while certain bigrams are prevalent across different sectors, their sentiment connotations can differ by industry. For example, 'collect data' is labelled as positive in the Health Care (HC) sector but negative in the IT industry. This can be interpreted as positive for HC as data collection is important for advancing medical research and improving patient outcomes. Conversely, in the IT sector, 'collect data' may carry negative connotations due to concerns over privacy and the potential misuse of personal information. This disparity highlights the limitations of a generic sentiment analysis dictionary, which would misclassify these terms and emphasises the importance of creating industry-specific dictionaries for accurate interpretation within each sector's unique context.

There is an ongoing debate in academic literature about whether human-based methods or machine learning (ML) techniques are more effective in creating sentiment dictionaries. Loughran and McDonald (2011) (LM) have posited that human judgment in selecting words is preferred to the "black box" nature of computer algorithms that have access to millions of data points throughout history. Consistent with Garcia et al.'s (2023) study, our research, using out-of-sample data, indicates that ML algorithms outperform the LM dictionary. In a comparative analysis of the ML dictionaries against the LM (human-based) dictionary, we observe that ML outperforms LM by an annualised alpha of 1.54% across all sectors. This finding reinforces the potential of ML in lexicon-based sentiment analysis and contributes to the ongoing debate between 'humans vs machines'.

The remaining sections of this thesis are structured as follows: The next section offers a literature review of the background of lexicon-based sentiment analysis. The second section presents the earnings call data and the progression to a sound dataset for analysis. The third section details the methods used to create the sentiment dictionaries. Finally, the fourth section examines the outcomes of the comparative analysis ('horserace') and discusses the new dictionaries created.

1.4. Literature review

Recent research in finance has increasingly used machine learning (ML) to understand and analyse the language in earnings calls, news media, and financial reports. Reviewing this literature helps us understand the advantages and potential challenges of using ML for textual analysis, guiding our approach to analysing earnings call Q&A segments. For example, Liang et al. (2021) applied ML techniques to earnings call texts, studying the post-earnings drift. Their research demonstrates ML's power in extracting insights from complex language. Moreover, Matsumoto et al. (2011) noted that earnings calls, especially their discussion periods, contain heightened information. This finding supports our decision to focus on the Q&A parts of these calls for applying ML techniques.

The application of machine learning in forecast stock prices through sentiment analysis is a rapidly evolving field, offering promising opportunities but also inherent challenges. As Loughran and McDonald (2016) warn, textual analysis is complex and context-dependent, leading to potential errors and systematic misunderstandings. The nuances of language, sentiment, and context in financial disclosures can result in misclassification errors. This could

affect the ML model's ability to accurately classify and interpret financial data, hence underlining the need for a cautious and prudent interpretation of the results. Our research overcomes limited semantic understanding by incorporating bigrams to capture the semantics of phrases. This will prevent our model from interpreting phrases such as ‘strong headwinds’ as positive statements when considered in isolation (unigrams). By adopting these higher-order linguistic structures, we improve the accuracy and context awareness of our model, leading to more precise and reliable dictionaries.

The debate on whether to use human-derived word lists or computational methods for sentiment analysis is still a central concern in financial textual analysis (Loughran and McDonald, 2020). While human-derived dictionaries, where we can visibly inspect the words and their scores, provide transparency and specificity, they involve subjective interpretation, introducing a potential bias. In contrast, ML approaches are capable of handling larger volumes of data and reducing subjective biases, however, face challenges related to data quality, feature selection, and model interpretability. Loughran and McDonald (2020) warn against the "black-box" nature of machine learning methodologies, where users are unaware of why a model reached a certain conclusion. Although these techniques boast significant advantages in processing power and objectivity, they may contain potential inaccuracies and have a lack of transparency.

Garcia et al. (2023) contribute to the ongoing debate between human-derived word lists and computational methods in sentiment analysis, demonstrating that machine-learning algorithms can significantly outperform existing human-curated dictionaries. Building on Garcia et al.’s (2023) findings, our study explores the potential for further refining these dictionaries by incorporating industry classifications. Words and phrases in financial discourse often show industry-specific characteristics, underlining the importance of context in textual analysis. Research by Hoberg and Phillips (2016) shows that firms within the same industry tend to use clustered vocabulary. This variation in language use across different industries signifies the value of developing specialised industry-specific dictionaries for effective sentiment assessment. Such findings encourage our hypothesis: By segmenting dictionaries according to industry to capture unique jargon, we aim to provide more precise and accurate sentiment analysis for each sector.

Recent research indicates an emerging trend where firms are modifying their language and sentiment in disclosures to accommodate machine readers, resulting in a reduction in

negative sentiment (Cao et al., 2023). This strategic adjustment presents a challenge for our machine learning model in predicting stock price movements. This dynamic could be likened to a cat-and-mouse game between models and firms. As companies and management teams evolve their vocabulary, this requires updates in the ML model data to represent these changes in the sentiment dictionaries. This suggests that sentiment dictionaries are continually evolving, meaning regular refining and updating of the dictionary is necessary to capture temporal fluctuations in positive and negative language over time.

Further advancements in natural language processing (NLP) techniques, especially those using deep learning like ChatGPT, present an intriguing alternative to traditional NLP approaches in predicting stock price movements. Huang et al. (2023) draw insights into the potential of large language models (LLM) to surpass traditional algorithms and ML techniques for financial text sentiment analysis. They showcase the superior performance of LLMs over the Loughran and McDonald (LM) dictionary and machine learning methods like support vector machines, random forest, and convolutional neural networks in extracting sentiment from financial texts. An example of a pre-trained language model that highlighted superior sentiment analysis is FinBERT (based on BERT) (Araci, 2019). LLMs like FinBERT can be fine-tuned using a variety of financial documents such as corporate filings, analyst reports, and earnings call transcripts. We posit that LLM models like FinBERT, fine-tuned on specialised financial data like the sentiment dictionary developed in this study, could surpass previous sentiment analysis methods.

The growing capabilities of these advanced language models in financial forecasting also come with limitations. Lopez-Lira and Tang (2023) find that sentiment analysis of news headlines using ChatGPT outperforms traditional NLP methods in predicting stock market returns. However, a limitation of this study is that the model was significantly concentrated on smaller stocks and firms with bad news, indicating that the predictive power might not work equally for all types of stocks. This informs our study, which encompasses a broader spectrum of companies, avoiding overemphasis on smaller firms with predominately negative news. Moreover, due to the lack of human control, ethical concerns are associated with deep learning methods, such as potential biases, risks of disinformation, manipulation, privacy, and regulatory considerations (Zaremba & Demir, 2023).

These findings indicate the potential superiority of LLMs in sentiment analysis, however, their effectiveness is heavily dependent on the training data. While Pre-trained LLMs

are powerful in prediction, they fall short in interpretability. In contrast, our ML approach offers powerful predictability due to the sophisticated approach to analysing textual data while also providing transparency by allowing visibility with a clearly defined list of positive and negative bigrams. We propose that integrating ML-developed sentiment dictionaries could further fine-tune LLMs like FinBERT, potentially leading to even better stock return predictions. By training our model on finance-specific data, our model will be more specialised than typical LLMs, which employ a broad range of texts for sentiment analysis.

Our research approach is informed by the findings of Darapaneni et al. (2022), who explored the application of sentiment analysis and deep learning in predicting stock price movements. Their research points to the need for clean data and appropriate regression model selections. This complements Jegadeesh and Wu (2013) study, which analysed the tone of 10-K reports based on market reactions, thereby providing an objective measure of document tone without subjective word classification. Their method shows the value of incorporating market reactions into language-based analysis. With these considerations, our supervised machine learning algorithm creates the dictionaries using stock price movements as a label.

Unlike studies such as Garcia et al. (2023) and Jegadeesh and Wu (2013), which use the entire 10-K statements to create sentiment dictionaries, our study focuses explicitly on the Q&A sections of earnings calls. These sections are valuable because they involve direct conversations between the company management team and financial analysts. The presentation sections are typically pre-prepared and scripted, making the Q&A section notably more informative due to fewer constraints on management's communication. This means the Q&A segment is rich in meaningful information, not overshadowed by irrelevant data or noise (Matsumoto et al., 2011). Research, such as the study by Hu et al. (2021), supports this selection by demonstrating the significant impact of Q&A sections on stock returns, highlighting their value as a source of 'price-sensitive' information. This segment discloses strategic information, often influencing stock prices significantly due to its immediacy and relevance. Our research focuses exclusively on the Q&A sections, as they have the greatest influence on stock price, and this provides meaningful data for machine learning (ML) applications.

We take inspiration from Bachhety et al. (2018) and Antweiler and Frank (2006) who use the Multinomial Naive Bayes classifier to analyse unstructured text data for predicting economic events. Their research provides valuable insights into effectively tackling the high dimensionality of text data, a challenge known as the "curse of dimensionality". The Naive

Bayes classifier, particularly the Multinomial variant, is used due to its strong text classification performance, driven by its ability to manage large-dimensional datasets. The Multinomial Naive Bayes classifier is chosen for its strong text classification capabilities and effectively manages sparse datasets. Furthermore, Gentzkow et al. (2019) highlight the significance of feature selection techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), to effectively filter common and rare words. This technique is instrumental in our study for extracting relevant terms. Integrating these methods guides the construction of our dictionaries, ensuring effective approaches to accurate sentiment analysis.

2. Data

Our study focuses on developing industry-specific dictionaries by analysing earnings call transcripts. We use stock price reactions as indicators, or labels, to guide our analysis. This approach allows us to identify impactful bigrams, which are then scored using our supervised machine-learning algorithm. Our research's success relies on analysing a clean dataset. In the following section, we describe the data used in our study and explain how we prepare and organise this data using Natural Language Processing (NLP) techniques.

2.1. Distribution of Earnings Call Transcripts

Transcripts were compiled from Thomson Reuters StreetEvents. StreetEvents curates corporate disclosures, financial documents, events and company updates, including corporate transcripts. We note that each company earning calls devote varying amounts of allocated time to the Q&A section and, therefore, will differ in size, with the average word count being 3747. Our dataset consists of quarterly earnings call transcripts from U.S. companies listed on major stock exchanges: 41% from the NYSE, 3% from AMEX, and 56% from NASDAQ. Initially, we had 144,505 transcripts. After merging this data with our mapping file, which contains key details such as the date and time of the calls, ticker symbols, and company names, the number of usable transcripts was reduced to 92,763. This reduction occurred due to mismatches between the entries in the transcript data and the mapping file, leading to the exclusion of incomplete or non-matching transcripts.

2.2. Speaker Identification and Data Extraction

The earnings calls data are provided as .xml file format. Earnings call transcripts are parsed to determine the speaker, whether it is the manager, analyst, or operator speaking.

Operator comments and questions from analysts are subsequently filtered out. The earnings calls .xml files have two main sections: the Management Discussion (MD) - a scripted presentation, and a Questions and Answers (Q&A) section where analysts ask questions to the management team. Our focus is on the Q&A section. Within the Q&A, we distinguish between the management's responses (answers) and the analysts' inquiries (questions). To achieve this, we label a speaker's contribution as an 'answer' if: 1) the speaker identifier can be mapped as an organiser or 2) the speaker spoke during the MD section of the call. All other contributions are labelled as questions.

For parsing .xml earnings call transcripts, we follow these steps:

1. Extract the Q&A section by finding section of the transcript that is tagged “q-and-a”.
2. Documenting questions by the same analyst speaker until interrupted by the operator.
3. Documenting answers by the same management speaker until a new speaker who also answers questions interrupts.
4. Requiring that a question must come after the operator speaks.

At the end of this process we create a dataframe, where we can link questions from analysts and answers from the management team, with their related text in the transcript. In our research, we specifically focus on analysing the responses from the management, as these are more impactful on stock price movements. This targeted approach assists our supervised machine learning algorithm in developing a sentiment dictionary that is effective in predicting stock price movements. Subsequently, while we label the analyst queries as questions, these portions of the text are ignored in our analysis.

2.3. Data Pre-processing

In the pre-processing phase, we clean textual data using standard NLP procedures. Punctuations are identified and removed, followed by tokenization of words. Next, all tokens underwent normalization: they are converted to lowercase and lemmatized. We exclude English stop words, single-letter tokens, numbers, and terms related to geography and colloquialisms, ensuring a coherent and analytically sound dataset. Each step contributes to creating an analytically sound dataset, ensuring our machine-learning algorithm can effectively learn and predict based on the cleaned and processed data.

To ensure robustness in our analysis, we applied a filter for textual depth, excluding transcripts with fewer than 100 words in the management team's responses (answers only). This was done to ensure our analysis would be based on sufficient and informative data, allowing for the generation of valid bigrams for our dictionary construction. After all relevant merges and data cleaning, our final sample size for analysis was 63,345 transcripts, associated with 5,661 unique firms. When focusing on the distribution of total transcripts per industry, we have the following: Information Technology (9,809), Consumer Discretionary (9,055), Health Care (8,295), Industrials (8,176), Financials (6,279), Energy (3,673), Materials (2,638), Real Estate (2,531), Consumer Staples (2,494), Communication Services (2,267), Utilities (1,367).

2.4. Financial and Stock Price Data

We require that the firms under analysis be matched to the Centre for Research in Security Prices (CRSP) and Compustat, where we source our data for company pricing data to calculate returns, and the relevant company industry (GICS), respectfully. CRSP provides daily stock returns, and we implement a rolling four-day window to capture stock price reactions post-earnings calls. These returns serve as labels for our methodology, subsequently helping formulate positive and negative bigram dictionaries. Compustat extracts relevant industry classifiers (like GICS codes), clustering firms into pertinent industry categories and ensuring that the developed dictionaries are finely tuned to sector-specific terminologies and jargon.

To calculate the stock price returns of companies during the earnings call period, we merge the price data with the transcript data. This merge was based on the company's ticker symbol and the date of the earnings call. This new dataframe not only includes all the information from the transcripts but also the opening stock price of the company on the day of the earnings call and the opening price four days after the call. These stock prices are used to calculate the returns, based on the percentage increase (decrease) from the first day opening price, to the fourth day opening price.

The motivation to focus on a four-day rolling window, we follow Garcia et al. (2023), where their study states that earnings calls price reactions are lagged due to a large portion of these earnings calls being in the afternoon, and the price reaction only being present in the following days. Heston and Sinha (2017) have shown that news sentiment can predict stock price movements for up to one or two days, highlighting the time-sensitive relationship between market sentiment and financial market outcomes. These findings, combined with the approach

of Garcia et al. (2023) show that our analysis should focus on predicting the stock market reaction on a short-term basis.

To validate the significance of the four-day rolling window applied around earnings calls, we analysed stock volatility during this period. Volatility is calculated as the standard deviation of returns. We measured the volatility in the four-day window post-earnings call and compared it to the average volatility over the following year, intentionally excluding the four-day window from our year-long observation. In our dataset, there is heightened volatility for companies around earnings calls periods compared to non-earnings call periods. Our analysis shows that stock volatility during these four days around an earnings call was, on average, 15.57% higher than periods outside of earnings calls. The heightened volatility during earnings calls highlights their importance, suggesting room for further analysis.

In addition to analysing company returns, we also extracted sector-specific returns from industry ETFs through CRSP. Typically, excess returns are calculated using broad market indexes like the S&P 500. However, our approach contrasts this norm by focusing on sector-specific ETFs to compute Cumulative Abnormal Returns (CARs). Our study calculates CARs as the stocks' four-day total return minus the respective industry's ETF four-day total return. This step helps control for broader market influences, allowing for more precise labelling of stock price reactions after earnings calls. This method aligns with Griffin's (2003) approach to computing absolute excess returns around the time of earnings calls. Our analysis included a comparison with excess returns calculated using the S&P 500, revealing that the difference between these two methods is negligible. Therefore, we specifically chose industry ETFs[1] as a control over broader market returns because they offer a more accurate capture of industry-specific excess returns. This allows for more targeted and precise industry dictionaries.

[1] The sector ETFs that were used are as follows:

- XTL: Communication Services
- XLB: Materials
- XLV: Health Care
- XLP: Consumer Staples
- XLY: Consumer Discretionary
- XLE: Energy
- XLF: Financials
- XLI: Industrials
- XLK: Information Technology
- XLU: Utilities
- IYR: Real Estate

2.5. Final Dataframe

After merging stock price data with the cleaned transcripts, this process results in a final dataframe, where each row represents an individual earnings call transcript. Key columns in this dataframe include 'eventid,' a unique identifier for each transcript; 'processed text,' which compiles all management responses from a call; the company's 'ticker' symbol; the 'Industry' classification; and the 'Excess industry return.' This dataframe, encompassing 63,345 transcripts, forms the foundation for all subsequent empirical analysis in our research.

2.6. Measuring Sentiment

This section describes our main textual analysis tools, Term Frequency-Inverse Document Frequency (TF-IDF) and the Multinomial Naive Bayes classifier (MNB). These two methods together score the bigrams that construct the dictionaries created. The scores generated by these methods are fundamental to constructing our sentiment dictionaries. A clear understanding of TF-IDF and MNB is essential, as they are instrumental in evaluating the bigrams used in our empirical analysis. The following sections will delve into each method, explaining how they contribute to the scoring process.

2.6.1. Term Frequency-Inverse Document Frequency (TF-IDF) in Text Analysis

To ensure the selected bigrams are suitable for analysis and do not contribute to the noise, it is crucial to filter them based on their relevance. This is achieved using the TF-IDF statistical metric:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (1)$$

Where:

- $tf(t,d) = \frac{occurences\ of\ term\ t\ in\ document\ d}{total\ number\ of\ terms\ in\ document\ d}$
- $idf(t,D) = \ln\left(\frac{total\ number\ of\ documents\ in\ corpus\ D}{number\ of\ documents\ containing\ term\ t}\right)$

TF-IDF effectively quantifies word relevance in two steps:

1. Term Frequency (TF): Represents how frequently a word appears in a specific transcript.

2. Inverse Document Frequency (IDF): Measures the rarity of a word across the entire corpus.

Together, they effectively assess word relevance, assigning higher scores to words that occur often in a document, but are rare across the corpus. For example, a common word like 'the' may have a high frequency in a single document (high 'TF'), but since it appears in almost every document, its 'IDF' score is low, leading to an overall low TF-IDF score. As a result, such common words are often disregarded in our analysis.

TF-IDF is beneficial for sentiment analysis because it can manage large amounts of text data. It identifies words and phrases within a text and scores these bigrams based on importance. This vectorisation process transforms unstructured textual data into a numerical format, creating a quantitative representation of the bigrams. Higher TF-IDF scores indicate that bigrams are more relevant, helping our machine-learning model create sentiment dictionaries. Its computational efficiency makes it an ideal choice for processing big datasets, ensuring that only meaningful bigrams are used for analysis.

To limit the dimensionality of our model, we will only include tokens with a TF-IDF score over a certain threshold. Our analysis suggests that 50,000 bigrams is the optimal threshold for training our model. This aligns with Garcia et al.'s (2023) research that uses 65,000 bigrams to train their model. These selected bigrams, transformed into vectors, are used as input (features) for the Multinomial Naïve Bayes Classifier model. The 'target' variable is the movement of the corresponding stock price within four days following the earnings call. In our study, train and test data are randomly split at a ratio of 80:20, with a seed for reproducibility.

2.6.2. Multinomial Naive Bayes Classifier (MNB)

After the TF-IDF process has identified the most relevant 50,000 bigrams from the earnings calls, these bigrams, along with their respective TF-IDF scores, are then inputted into our supervised machine learning model – the Multinomial Naïve Bayes Classifier (MNB). When paired with TF-IDF, the model captures the relationship between bigrams and stock price movements. MNB is used due to its strong text classification performance, driven by its ability to manage large-dimensional datasets. The Multinomial Naïve Bayes classifier is a supervised machine learning algorithm specialising in classification tasks, such as text classification. At its core, the classifier operates on Bayes' Theorem, using conditional probability. This method calculates the likelihood of an event - such as a bigram's score, based on past events - like stock

price movements, under the principle of independent predictors. For instance, when analysing a transcript, the classifier assigns a score (feature importance) to a specific bigram based on its probability to influence stock price movement. In simpler terms, a higher absolute score implies the bigram's stronger potential impact on stock price direction.

The unique aspect of this classifier is its use of a multinomial distribution, suitable for modelling word occurrences. To illustrate, the Naïve Bayes Classifier will determine the sentiment score of a certain bigram 'x' based on its association with stock-price movement in the training data. The assumption of conditional independence means that the sentiment score calculated on bigram 'x' is calculated without consideration of other bigrams within the same document. Even if these bigrams depend on each other, the independent consideration for each bigram is why it is known as 'Naïve'. This may seem oversimplified, especially considering the interconnectedness and semantic relationships between words in financial transcripts. However, this simplification allows for model efficiency when handling high-dimensional datasets and is widely known for its effectiveness in classifying documents in practical, real-world applications.

2.6.3. Other Tested Models

Despite the efficiency and effectiveness of MNB, it is worth acknowledging that other models were considered. For instance, models like Ridge regression, Lasso regression, and Support Vector Machines (SVM) have been popular in similar tasks.

However, these models do not perform effectively in our specific context. Ridge and Lasso regression, being modifications of linear models, might have struggled to capture complex non-linear relationships present in the high-dimensional textual data. Such models can often become restrictive when working with a vast set of features that might exhibit intricate relationships. Likewise, SVM might have faced challenges due to the high dimensionality and sparsity of data.

In contrast, the Naïve Bayes classifier offers a more straightforward and efficient approach, inherently resilient to irrelevant features, which are commonplace in high-dimensional textual datasets. However, one limitation of this binary classifier approach is its requirement for categorical data, which restricts its capacity to capture the magnitude of abnormal returns. Since MNB needs labels to be either positive or negative, it treats both a 5% and a 0.1% stock increase identically when scoring the sentiment of bigrams.

2.6.4. Hurdle Distributed Multinomial Regression (HDMR)

HDMR as proposed in Text Selection by Kelly et al. (2019), offers an intriguing alternative to the Multinomial Naïve Bayes classifier approach. HDMR is a high-dimensional regression technique that finds intricate relationships within extensive textual datasets. One advantage of HDMR is its ability to capture abnormal returns effectively, providing a more detailed understanding of financial sentiment. Unlike the binary classification approach of MNB, HDMR allows for a more nuanced analysis by capturing the loadings of each dictionary term and enabling the comparison of the impact of individual terms. For example, bigrams from an earnings call associated with a stock price increase of 5% are assigned higher importance and score in the HDMR model compared to a 0.1% increase. This added complexity enhances its capability to identify significant words or phrases in earnings calls that are associated with price movements. As a result, HDMR presents itself as a promising alternative approach for future research in financial text analysis, offering the potential for deeper insights and improved predictive accuracy.

2.7. Implementation of TF-IDF and MNB

In this section we describe the implementation of our textual analysis tools, including the parameters and assumptions used. To execute our analyses with the Multinomial Naïve Bayes classifier, the data is split into training and test sets, allocating 20% to the out-of-sample data. The Naïve Bayes is characterised by its difficulty reproducing results (Loughran and McDonald, 2016). Therefore, A consistent random state is applied, ensuring the reproducibility of our results is possible.

The model is trained with a grid search over possible alpha values [0.01, 0.1, 0.5, 0.75, 1, 2, 5, 10], where ‘alpha’ represents the optimisation parameter of the model. This helped in pinpointing the configuration that maximised the model's performance. The Naïve Bayes Classifier model is trained to identify 50,000 bigrams. Each bigram is assigned a sentiment score by subtracting the log probabilities for the positive and negative classes. The highest scores indicate the most positive bigrams, while the lowest scores correspond to the most negative ones.

These scores are used to construct the dictionary, where a score over zero resulted in the bigram being assigned to the positive dictionary and a score less than zero being assigned to the negative dictionary. This results in a dictionary of 25,086 positively labelled bigrams and

24,914 negatively labelled bigrams. The model is better at classifying negative bigrams, achieving higher precision and recall for the negative class than the positive class. This means that the model more accurately identifies actual negative bigrams (precision) and captures most of them present in the data (recall). This results in our model being more effective in a shorting strategy.

2.8. Construction of Dictionaries

In our study, we develop a 'Universal Dictionary,' comparable to existing dictionaries in the literature, trained on all data without industry-specific considerations. Additionally, we introduce 'Industry-Specific Dictionaries,' each tailored and trained exclusively on data from one of the 11 GICS sectors. The process of formulating these dictionaries can be represented as two phases, and is captured in the following modelling equations:

Phase 1: Universal Dictionary

$$\underset{excess\ returns}{Y} \approx \underset{transcript\ text}{f(x)} \qquad (2)$$

Where:

- f: Multinomial Naïve Bayes classifier trained on all stocks and transcripts.

Phase 2: Industry-Specific Dictionaries

$$\underset{excess\ returns}{Y_j} \approx \underset{transcript\ text}{f_j(x_j)} \qquad (3)$$

Where:

- j: represents a specific industry.
- f: Multinomial Naïve Bayes classifier trained specifically on the j industry.
- x_j: Transcript text corresponding to companies in the j industry.

When splitting up dictionaries tailored to specific industries, we aim to capture the variations in linguistic usage across sectors. We hypothesise that certain words and phrases will carry different connotations based on their industry context. Should our hypothesis hold,

industry-specific dictionaries will likely outperform a universal dictionary within their corresponding sectors.

We anticipate that industry-specific dictionaries will be less accurate when applied outside their respective industries. For example, a dictionary tailored to the Financials sector may be less efficient when applied to the Health Care industry.

By investigating these cross-industry differences, we aim to show the industry-specific specialised language and its impact on stock price predictions, thus refining our sentiment analysis model.

3. Empirical Design

In this section, we present the empirical design of our study, which focuses on comparing our newly constructed dictionaries with other established dictionaries in the field. Using out-of-sample data, we create a trading strategy to evaluate each dictionary's predictive capabilities. Our design has two main parts: first, using each dictionary to score the sentiment of these out-of-sample earnings calls; second, using these sentiment scores to formulate a trading strategy to test their predictive power. The trading strategy tests these dictionaries' ability to forecast stock price movements.

We engage in a horserace between our dictionaries created with our machine-learning methods against established dictionaries, such as Loughran and McDonald (2011) (LM), and Garcia et al. (2023) (GHR). We also benchmark our dictionaries against a traditional pre-trained sentiment analysis model. This model is trained on a wide variety of texts and is good at understanding general language but is not trained on finance exclusive data. Throughout this study, we will refer to the general-purpose sentiment model (not finance-specific) as 'traditional sentiment analysis'. By comparing our approach against this model, we can evaluate if the finance-specific language of earnings calls (and the dictionaries we have built to understand this language) can yield more accurate stock price predictions than general language models.

3.1. Creating Sentiment Scores

The foundation of our approach is a proof-of-concept trading strategy derived from sentiment scores of transcripts. Using bigrams classified as positive or negative, we compute the sentiment for each out-of-sample transcript.

The trading strategy is performed on the 11,936 out of sample transcripts (20%) using Equation (4):

$$sentiment = \frac{\#Positive\ Bigrams - \#Negative\ Bigrams}{\#Bigrams\ in\ Transcripts} \quad (4)$$

By comparing the count of positive bigrams against the count of negative ones and normalising it with the total bigrams in a transcript, we get a sentiment score. This exercise creates a dataframe assigning each out-of-sample transcript with its corresponding sentiment score.

This is consistent with previous studies in the literature, (Loughran and Macdonald 2011; Garcia et al., 2023), that employ a traditional 'bag-of-words' approach to compute sentiment scores for documents. Specifically, these studies calculate the sentiment by summing up the frequencies of terms from either the positive or negative dictionary present in the document. This sum is normalised by dividing it by the total word count of the document. This method is rooted in long-standing practices where sentiment scores are deduced from the sum of term frequencies of specific dictionary members, scaled by the document's size.

3.2. Trading Strategy

To evaluate out-of-sample predictive performance in economic terms, we design a trading strategy that leverages sentiment estimates for prediction. A sentiment-based ranking is applied, sorting the out-of-sample transcripts from the most positive to the most negative, subsequently leading to the constitution of equally sized "long" and "short" portfolios. Before conducting a robustness test, we selected the top 25% of transcripts with the highest positive sentiment scores and the bottom 25% with the lowest sentiment scores. This trading strategy calculates portfolio returns as the CARs of the four-day excess returns using the CRSP data detailed in section 2.1. The top bracket was 'longed' while the bottom was 'shorted'. This creates a portfolio of 5,968 transcripts, or for interpretability, a long strategy of 2,984 companies, and a short strategy of 2,984 companies. In the trading strategy, we use equal weighting, which is characterised by its simple and robust means for assessing the predictive power of sentiment irrespective of firm size, as suggested by Ke et al. (2019). Therefore, in our strategy, our initial investment is equally distributed among all companies in the portfolio. When comparing our performance with existing dictionaries, we apply the same trading strategy for consistency.

4. Results

Section 4.1 studies the Universal Dictionary returns from the trading strategy (described in Section 3). We compare the performance of our machine-learning algorithm to that of Garcia et al. (2023) to see if our approach is valid. We also compare the machine learning sentiment analysis approach to the human-based LM approach, and traditional sentiment analysis (trained on non-finance specific data). Section 0 examines the performance of the 11 industry-specific dictionaries we created. We analyse the bigrams created and portfolio returns to answer the question of whether there is validity in creating industry-specific dictionaries.

4.1. Universal Dictionary Results

As seen in Figure 1, when using a machine learning model for sentiment analysis as a predictive tool for stock returns, the trading strategy yields positive returns irrespective of the portfolio size, however it is most effective when operating at sentiment extremes. This shows validity in our ML model's ability to capture sentiment, as more extreme sentiment scores are reflecting higher CARs. Another reason for the reduced returns observed with larger portfolio sizes, is that the management team will try to mask poor results with a positive tone, with the aim to deceive the audience. In Price et al. (2012) study, they find that in earnings calls, the tone is over five times more positive than negative. This means many transcripts, where sentiment should be slightly negative, might be incorrectly scored as positive. This contributes to the noise for transcripts where sentiment is not easily detected and is one reason why sentiment analysis is more accurate at the extremes.

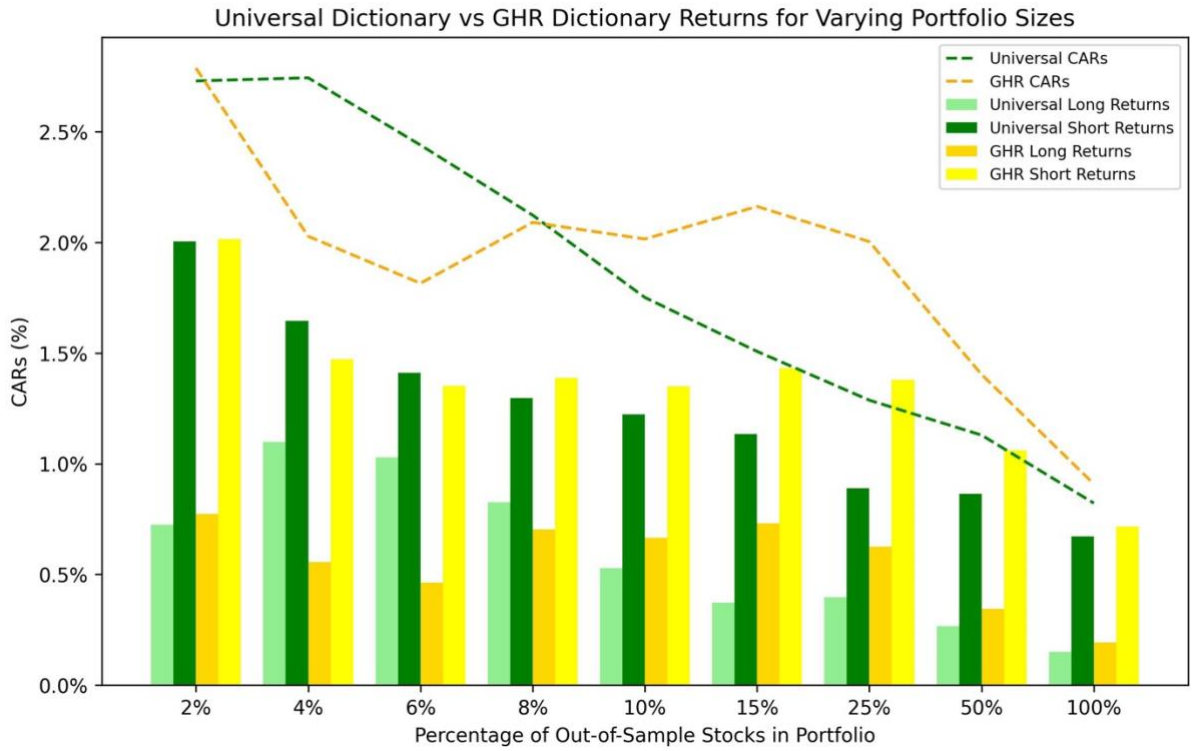


Figure 1: Universal vs GHR Dictionary: Performance Across Portfolio Sizes

The figure is designed as a robustness test for ML dictionaries across different proportions of out-of-sample data (11,936 transcripts). In our approach, 2% represents a portfolio of 240 companies and only includes the most positive and negative transcripts, whereas 100% captures the whole out-of-sample data.

The machine learning models excel at detecting negative sentiment in earnings call transcripts despite management's efforts to appear positive. When examining the Universal and GHR dictionaries at the 8% portfolio size, the shorting strategy returns 1.35% on average, outperforming the 0.77% average return from the long strategy. This trend holds, with short-return strategy outperforming long-returns across each tested portfolio size. This precision makes the model particularly useful for predicting stock price declines and favours short-selling tactics among traders.

In our comparison between our Universal dictionary and the GHR dictionary, we observe that our dictionary outperforms GHR's when detecting transcripts with extremely positive or negative sentiments (refer to Figure 1). On the other hand, GHR dictionary is better at correctly detecting sentiment for larger portfolio sizes.

Using a trading strategy with an out-of-sample size of 11,936:

- Our Universal Dictionary is more effective for portfolio sizes of 240-720, by an average total return difference of 1.28%.
- GHR's dictionary outperforms with larger portfolio sizes of 1,200-2,388, by an average total return difference of the two portfolios of 2.00%.

Both dictionaries intersect at a cumulative total return for a portfolio size of 960 (8%), delivering an average return of '2.11%'. Consequently, this robustness test informs our decision to set the portfolio size at 960 for the Universal Dictionary, ensuring an optimal balance between portfolio size and performance, thereby facilitating a fair comparison with other dictionaries in the next section.

Table 1 shows the portfolio returns for the Universal Dictionary, GHR Dictionary, and LM Dictionary. The returns from the two ML methods (Universal, GHR Dictionary) is compared to that of the human-based Dictionary (LM). Note that for short returns, a negative value indicates a gain, as the strategy profits from a decline in prices. Sharpe ratio and t-statistic are calculated using CARs.

Table 1: Comparing Portfolio Returns of Human Dictionaries vs Machine Learning Approaches

Dictionary	Long Return	Short Return	CARs	Sharpe Ratio	t-statistic
Universal Dictionary	0.83%	-1.30%	2.13%	0.34	10.15
GHR	0.70%	-1.39%	2.09%	0.31	9.59
LM	0.18%	-0.68%	0.86%	0.13	3.81

In this section, we aim to evaluate various sentiment analysis tools, applying each to the same trading strategy for comparison. The dictionaries being compared include the 'Universal Dictionary' developed in this study, GHR's machine learning (ML) approach, and the Loughran and McDonald (LM) 'human-based' approach. These dictionaries are tested on a portfolio comprising 960 out-of-sample stocks, divided equally between long (480) and short (480) positions.

Our findings reveal that the ML approaches (Universal, GHR dictionary), perform significantly better than the LM 'human-based' approach. The 'Universal Dictionary' created in this study, achieves an annualised alpha of 8.52%. When analysing the Sharpe ratio, a

comparison between ‘Universal Dictionary' and the GHR dictionary shows comparable efficiency, with scores of 0.3351 and 0.3144, respectively. Both dictionaries exceed with a t-statistic of 10.1529 for our dictionary and 9.5893 for GHR. Looking at the LM dictionary, the predictive power is far lower, with an annualised alpha of 3.44% and a t-statistic of 3.8108. This shows the outperformance of the ML-based approach in constructing dictionaries and contributes to the debate of 'Humans vs Machines'.

Table 2 shows traditional sentiment analysis results compared to the Universal Dictionary results across all industry. The table shows the number of out-of-sample transcripts identified as positive (negative) sentiment. This table highlights the importance of having a ML model trained on Finance-Specific data versus a general language model trained a variety of texts. Note that for short returns, a negative value indicates a gain, as the strategy profits from a decline in prices. Sharpe ratio and t-statistic are calculated using CARs.

Table 2: Traditional Sentiment vs Universal Dictionary Returns

	Traditional Sentiment		Universal Dictionary	
Industry	Number of Transcripts classified as Positive (Negative)	LR SR CARs	Number of Transcripts classified as Positive (Negative)	LR SR CARs
Information Technology	1,814 (1)	-0.08% -0.98% 0.90%	529 (1,257)	0.53% -2.05% 2.58%
Consumer Discretionary	1,717 (5)	0.25% -0.19% 0.44%	800 (901)	0.26% -0.69% 0.95%
Health Care	1,543 0	-0.39% 0.07% (0.46%)	319 (1,209)	0.35% -0.31% 0.66%
Industrials	1,585 (2)	-0.16% -0.56% 0.40%	562 (1,008)	0.13% -1.01% 1.14%
Financials	1,230 (2)	0.30% -0.19% 0.49%	577 (637)	0.47% -0.28% 0.75%
Energy	708 (2)	-0.99% -0.48% (0.51%)	161 (533)	-0.58% -1.76% 1.18%
Materials	516 0	0.04% -0.60% 0.64%	144 (368)	-0.05% 0.49% (0.54%)
Real Estate	501 (1)	-0.22% -0.56% 0.34%	191 (305)	0.38% -0.74% 1.12%
Consumer Staples	481 (2)	-0.56% -0.77% 0.21%	234 (243)	0.18% -0.89% 1.07%
Communication Services	418 (1)	-0.35% -0.32% (0.03%)	117 (295)	0.92% -0.36% 1.28%
Utilities	265 (1)	0.37% 0.08% 0.29%	42 (219)	-0.13% 0.87% (1.00%)
Average (industries)	979.82 (1.55)	-0.16% -0.41% 0.25%	334.18 (634.09)	0.22% -0.61% 0.84%

This section contrasts the ML dictionary created in this study, trained on finance-specific data, with general-purpose sentiment analysis models. As discussed in the literature review section, sentiment analysis in finance should utilise dictionaries trained on finance-specific data to capture the unique language accurately. Kelly et al. (2021) emphasise the advantage of dictionaries tailored to the context of the dataset being analysed, freeing researchers from relying on pre-existing, general-purpose sentiment dictionaries. Similarly, Consoli et al. (2022) show that their finance-specific sentiment analysis algorithm outperforms generalised models by focusing on topic-specific economic and financial reports.

Table 2 reinforces this by illustrating the superior performance of the ML model tailored to financial data, achieving a higher CARs in 82% of industries. As discussed, the management team may mask poor results with a positive tone in earnings calls. Our results show that traditional sentiment analysis overwhelmingly classifies earnings call transcripts as positive—on average, 980 out of 982 instances across all industries. This equates to a 99.8% rate of positive sentiment classification, indicating a significant misclassification bias within general-purpose models. This overestimation of positive sentiment suggests that generalised models fail to recognise and accurately interpret the specialised language used in finance. This also represents how the management team uses a positive tone to present the company favourably to the audience.

Our results reveal a bias: traditional sentiment analysis often predicted positive sentiment, yet this prediction did not align with actual positive returns. For traditional sentiment analysis, long positions averaged a loss of -0.16%, in comparison with our finance-specific ML model, which realised an average gain of 0.22%. This significant difference highlights the limitations of the traditional model's ability to discern genuinely positive sentiments, suggesting it might be overfitting by recognising positive words without context. Models trained on finance-specific data, like ours, excel in detecting sentiment from the nuanced language management use.

The objective of this section is to validate the effectiveness of our ML sentiment dictionary, shown in Figure 1 and Table 1. The results indicate the reliability of our approach, ensuring that our comparison of industry-specific dictionaries is adequate. Table 2 shows that the precision of sentiment analysis improves with the refinement of dictionaries to the specific linguistic context of finance. In the next section, our study focuses on developing industry-

specific dictionaries, enabling a more granular approach to the jargon and terminology of each sector.

4.2. Industry-Specific Dictionary Results

We analysed industry-specific outcomes by separating training and testing data for each sector. This allows for the construction of new industry-specific dictionaries. These industry-specific dictionaries will be tested against the 'Universal Dictionary' created from the whole dataset. We will also compare the results with GHR's ML method, LM 'human-based' approach, and the pre-existing traditional sentiment model. It is important to note that these sectors have varying levels of transcript data available: Information Technology (9,809), Consumer Discretionary (9,055), Health Care (8,295), Industrials (8,176), Financials (6,279), Energy (3,673), Materials (2,638), Real Estate (2,531), Consumer Staples (2,494), Communication Services (2,267), Utilities (1,367). Due to the small number of transcripts available for some industries, we placed a threshold of at least 130 stocks in the portfolio, which represents a portfolio size of 50% of out-of-sample data (25% long, 25% short). The 50% of out-of-sample data is used across all industries for consistency.

The main hypothesis of this study is that tailored ML dictionaries, refined by industry, will yield more precise sentiment analysis results. This is because ML thrives on precise, high-quality data. We expect that an industry-specific dictionary will more accurately reflect sentiment in its respective sector. For example, we find that the term 'Covid Related' is seen positively in the Health Care industry, but 'Impact Covid' holds negative sentiment for the Consumer discretionary industry. Due to industry semantics, a one-size-fits-all dictionary might incorrectly assess these terms across different industries.

Table 3 shows the returns for all 11 Industry-Specific Dictionaries created. The table shows the Long returns, Short Returns, CARs (L-S Returns), Sharpe Ratio, t-statistic, and volatility. Note that for short returns, a negative value indicates a gain, as the strategy profits from a decline in prices. Sharpe ratio and t-statistic are calculated using CARs.

Table 3: Industry-Specific Dictionary Performance

Industry-Specific Dictionary	LR SR CARs	Sharpe Ratio	t-statistic	Volatility
Information Technology	0.54% -1.62% 2.16%	0.28	4.43	7.73
Consumer Discretionary	0.42% -0.61% 1.03%	0.16	1.89	6.29
Health Care	0.15% -0.39% 0.54%	0.07	0.96	7.34
Industrials	-0.27% -0.38% 0.11%	0.02	1.17	5.82
Financials	0.28% 0.22% 0.06%	0.01	-0.88	4.20
Energy	-0.62% -1.58% 0.96%	0.16	3.30	6.13
Materials	0.42% -0.10% 0.52%	0.11	0.21	4.67
Real Estate	-0.41% 0.09% (0.50%)	-0.13	-0.26	3.71
Consumer Staples	-0.09% 0.03% (0.12%)	-0.02	-0.05	5.37
Communication Services	0.93% -1.16% 2.09%	0.33	1.96	6.36
Utilities	-0.22% -0.04% (0.18%)	-0.04	0.09	4.01
Average (industries)	0.10% -0.50% 0.61%	0.09	1.16	5.60

Across the 11 industries, our findings exhibit mixed results. In the analysis, we find that the Industry Specific Model has positive returns for the seven largest industries (number of transcripts in the dataset) and negative returns for three of the four smallest industries. We deduce that the poor results for smaller industries are due to a lack of data available, limiting our model's ability to create sector-specific dictionaries that hold strong predictive power. For example, the utility sector, with only 1,367 transcripts, leads to the dictionary being trained on a smaller subset of 1,093 transcripts, encompassing a total of 451,701 bigrams. Our model is required to create a dictionary of 50,000 bigrams from 451,701 bigrams available to select. This compares to our biggest industry, Information Technology, which has 2,124,754 bigrams, and the model selects 50,000 bigrams. This results in the model selecting less than 2.5% of the available bigrams for the IT sector, compared to over 11% for Utilities. This lack of data for certain industries results weaker performance.

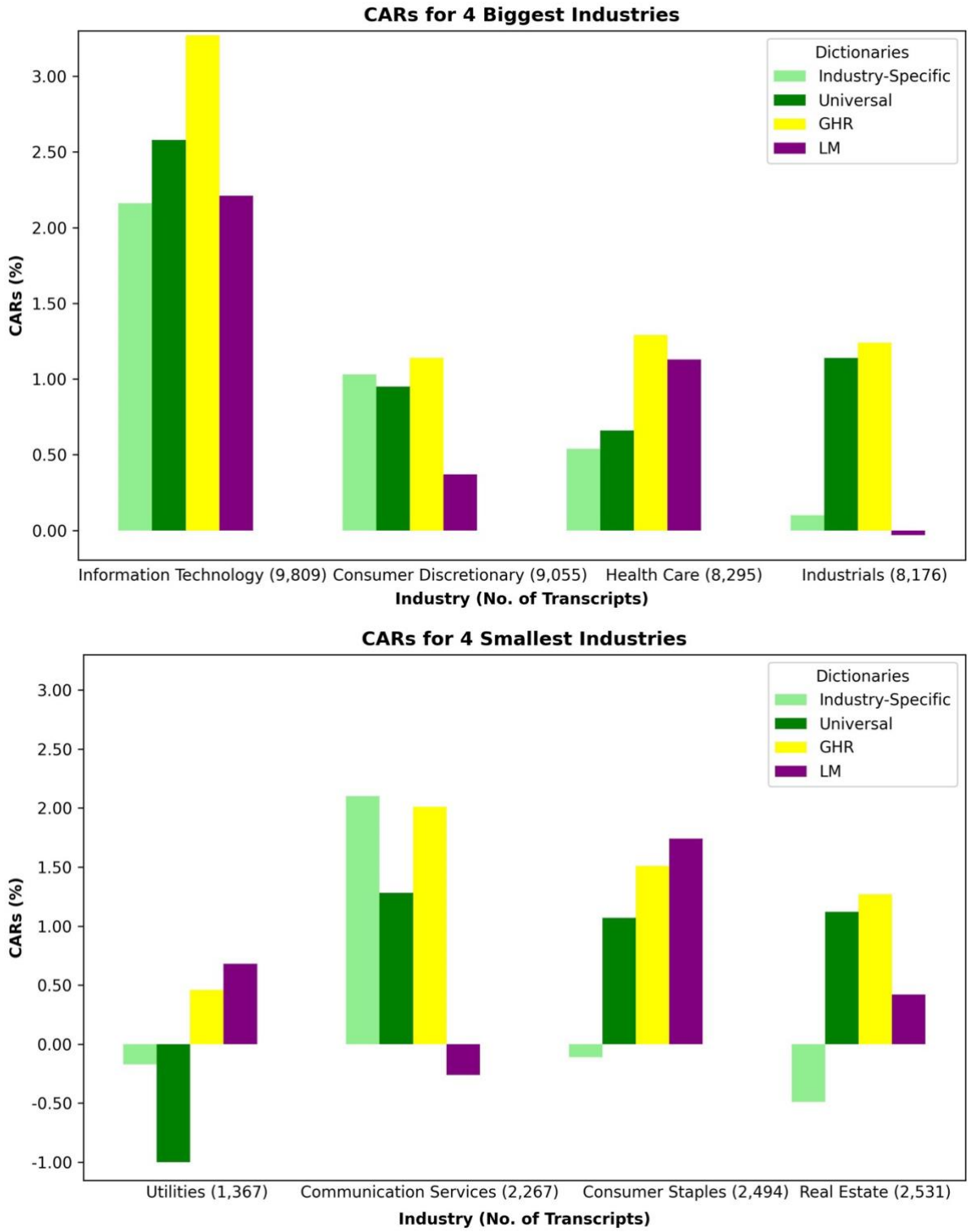


Figure 2: CARs for Four Biggest vs Four Smallest Industries

The graphs in Figure 2 show the CARs for the Industry-Specific, Universal, GHR, and LM Dictionary. This compares the four largest industries by number of transcripts versus the four smallest industries by number of transcripts. This shows the better performance for the Industry-Specific Dictionaries when trained on the four largest industries, compared to the poor results in the four smallest industries.

In comparing sectors with the most and least transcript data (refer to Figure 2), we observe that larger datasets correlate with better investment returns. The sectors with the most data—Information Technology, Consumer Discretionary, Health Care, and Industrials show better performance when using our industry-specific dictionary. On average, these sectors achieve a higher annualised return of 4.95%, a stronger t-statistic of 5.13, and a superior Sharpe ratio of 0.18. In contrast, the sectors with fewer transcripts—Utilities, Communication Services, Consumer Staples, and Real Estate—exhibit weaker performance. Only one of these sectors presents a positive trading strategy, showing that a lack of data affects the model. Their returns average 2.91%, with a t-statistic of 1.98 and a Sharpe ratio of 0.13.

Moreover, while the industry-specific dictionary aligns more closely with the Universal and GHR's dictionary for the four largest industries, the smallest industries lag significantly in performance compared to dictionaries backed by more extensive data. This disparity highlights a broader implication: the volume of data a dictionary is trained on directly influences its accuracy in sentiment analysis. It highlights the principle that a successful dictionary requires substantial training data to function effectively.

Table 4 shows the bigrams created in the top three industries by number of transcripts – Information Technology, Consumer Discretionary, and Health Care. The consistent bigram classification shows bigrams that are common in all three industries, as well as common in the Universal Dictionary. The table also shows where bigrams are found as positive in one industry but holds negative sentiment in another.

Table 4: Analysing Bigrams from the Top Three Industries by Number of Transcripts

Consistent Bigram Classification Across Three Industries		Bigrams Classified as Positive in One Industry and Negative in Another	
Positive	Negative	Positive	Negative
Increase Investment	Expect Significant	Covid Related (HC)	Impact Covid (CD)
Positively Impact	Impact Actually	See Disruption (IT)	See Disruption (CD)
Growth Objective	Continue Opportunistic	Increase Traffic (CD)	Increase Traffic (IT)
Saw Opportunity	Increase Expense	Collect Data (HC)	Collect Data (IT)
Great Opportunity	Development Cost	Technology Today (IT)	Technology Today (HC)

As hypothesised, analysing bigrams within industry-specific dictionaries reveals common and unique bigrams captured in both the universal (trained on all data) and industry-specific dictionaries. In examining the Information Technology (IT), Consumer Discretionary (CD), and Health Care (HC) sectors, we find common bigrams like 'increase investment'— positive across the board—and 'increase expense'— negative for the three industries. This example illustrates bigrams' effectiveness in providing context, where the term 'increase' is associated with both positive sentiment ('increase investment') and negative sentiment ('increase expense'). Bigrams, therefore, are crucial for providing clarity that single words (unigrams) lack. Their use in dictionary construction offers a refined approach over the unigram-based LM dictionary, enabling more precise sentiment analysis, as supported by Garcia et al. (2023).

While some bigrams' sentiment remains consistent across all three sectors, certain bigrams are classified as positive in one industry but negative in another. For instance, 'see disruption' carries positive sentiment in the IT industry, yet it is negative in the consumer discretionary dictionary. This is straightforward in the sense that 'see disruption' is positive for IT because it might entail a new technology that would be positive for the company, whereas 'see disruption' may be negative for Consumer Discretionary as it might suggest new entrants, technologies, or changes in consumer behaviour. This contradiction suggests that a one-size-fits-all dictionary can misinterpret these terms. Hence, an industry-specific dictionary is recommended for accurate sentiment analysis within each sector.

As seen in the results in this section, our study demonstrates that different industries contain specialised language. We show that industry-specific dictionaries more effectively capture each industry's unique bigrams and specialised language. We highlight that certain phrases carry different sentiments across industries — what is positive in one industry may be negative in another. This discrepancy leads to mislabelling words (tokens) in universal dictionaries. As a result, using industry-specific dictionaries enhances sentiment analysis for companies within their respective sectors.

Although we cannot show evidence through this study that on aggregate industry-specific dictionaries outperform traditional universal dictionaries trained on all sectors, we show a proof-of-concept for industry-specific dictionaries being useful in sentiment analysis. We observed that industries trained on a larger dataset excelled in performance and shows promising potential. Future research should utilise a sufficiently large and balanced dataset to conduct a robust comparison between industry-tailored and universal sentiment dictionaries.

4.3. Sentiment Analysis Applications

This section presents insights for using sentiment analysis in finance. We explore the strategies that yield the best results. We focus on reasons for creating dictionaries using bigrams over unigrams for more accurate sentiment detection. We also show that the best sectors for sentiment analysis contain volatility and further insights to optimise returns.

Table 5 shows words that are labelled as positive (negative) in the LM dictionary. We find how often these words when paired as a bigram appear positive (negative) in the Universal Dictionary. The table shows how words can be mislabelled when scored in isolation.

Table 5: Disambiguating Sentiment Between LM Unigrams Contained in Bigrams

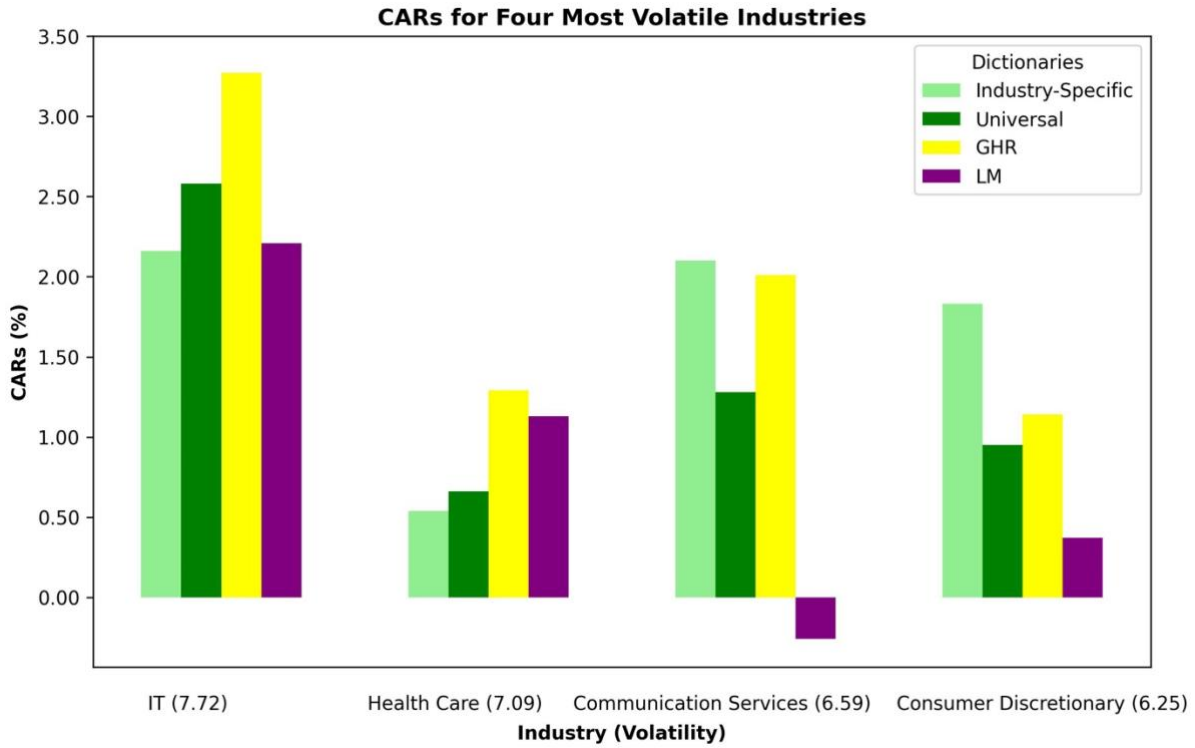
Positive Words (LM)				Negative Words (LM)			
Word	% Occurrence in Positive Bigrams	% Occurrence in Negative Bigrams	Negative Bigram Example	Word	% Occurrence in Positive Bigrams	% Occurrence in Negative Bigrams	Negative Bigram Example
Grow	73%	27%	'low growth'	Impact	39%	61%	'positive impact'
Improve	61%	39%	'slight improvement	Hard	32%	68%	'work hard'
Drive	69%	31%	'drive cost'	Tax	21%	89%	'tax benefit'
Profit	100%	0%	N/A	Believe	37%	63%	'believe best'
Results	61%	39%	'impact result'	Lower	33%	67%	'lower cost'
Good	68%	32%	'quite good'	Expect	45%	55%	'expect growth'
Increase	64%	36%	'cost increase'	Decline	20%	80%	'rate decline
Margin	66%	34%	'margin pressure'	Take	37%	63%	'take advantage'
Confident	55%	45%	'remain confident'	Debt	30%	70%	'reduce debt'
Cash	40%	60%	'cash burn'	Cost	43%	67%	'lower cost'
Continue	72%	38%	'continue evaluate'	Effect	44%	66%	'cost effective'

Our examination of unigrams in the Loughran and McDonald (LM) dictionary, detailed in Table 5, reveals limitations when compared to bigrams. This is due to the unigrams being labelled as strictly positive or negative, no matter the context. However, when these words are paired with others and captured as bigrams, their sentiment can change.

For example, the LM dictionary categorises words like 'increase' as positive in isolation. However, when we analyse these words within the context of bigrams in our dictionary, their sentiment shifts. For example, 'increase' appears positive in 64% of bigrams but negative in 36%, an example being 'cost increase.' This difference highlights the importance of context in sentiment analysis.

Similarly, words such as 'grow' and 'good' are seen as positive in the LM dictionary, but in our bigram dictionary, they appear negative 27% and 32% of the time, respectively, in contexts like 'low growth' and 'quite good.' On the other hand, some words traditionally viewed negatively, like 'tax' and 'debt,' show a positive sentiment in certain bigrams, such as 'tax benefit' and 'reduce debt,' respectively.

Our findings capture the limitations of unigram-based sentiment dictionaries. By including bigrams in dictionaries, the practitioner can more accurately reflect the context and true sentiment of phrases, avoiding the mislabelling that often occurs with unigram-only approaches. Our findings align with those of Garcia et al. (2023), emphasising the necessity for sentiment analysis tools to consider the full linguistic context. This approach enhances the precision and reliability of financial sentiment analysis, as bigrams allow for a more nuanced and dynamic classification compared to unigrams.



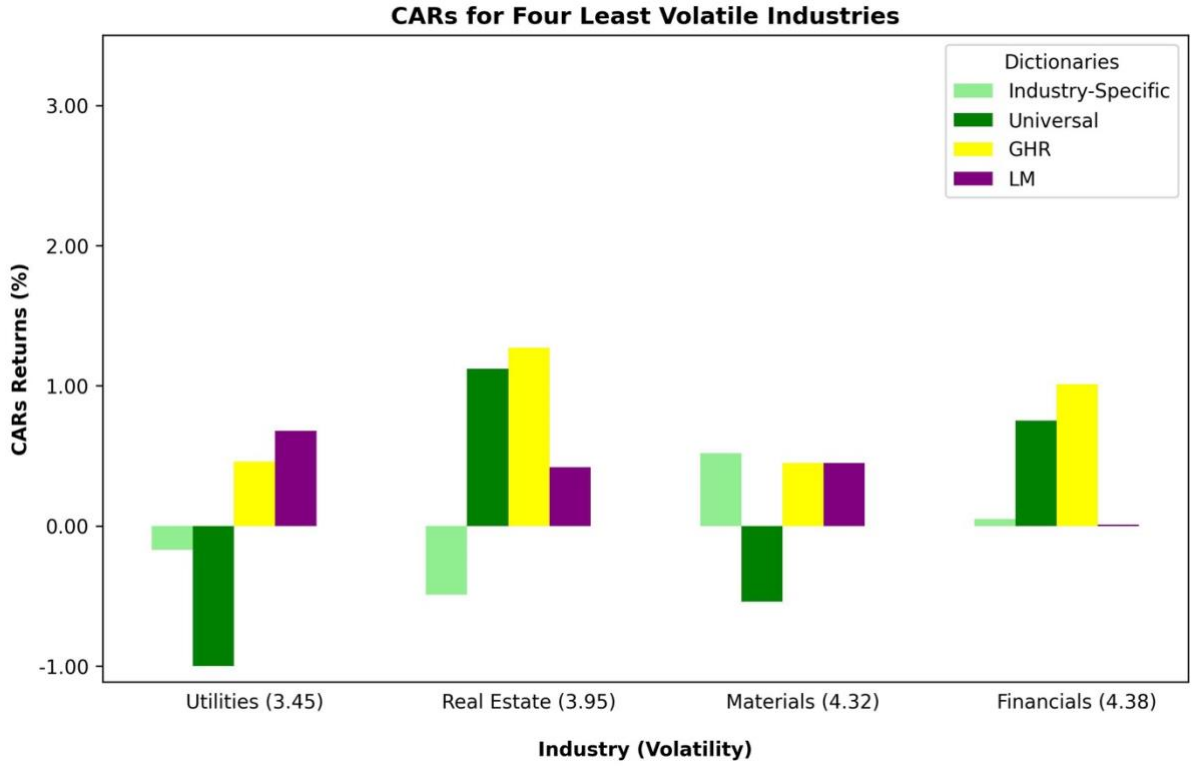


Figure 3: CARs in Four Most Volatile vs Four Least Volatile Industries

The graphs in Figure 3 show the CARs for the Industry-Specific, Universal, GHR, and LM Dictionary. This compares the four most volatile industries versus the four least volatile industries.

Higher stock volatility represents more uncertainty about future stock prices, presenting an opportunity for informed investors to gain higher returns. However, uncertainty also implies that stock-related news is harder to interpret. Despite this, ML-based approaches exhibit strong accuracy in predicting stock price movements.

In comparing sectors with the most and least volatility (refer to Figure 3), we observe that more volatility exhibits greater investment returns. Information Technology, Health Care, Communication Services, and Consumer Discretionary display more volatility and demonstrate statistically significant performance and a better risk-adjusted return, evidenced by the Sharpe ratio. These sectors, on average, achieve a higher annualised return of 5.62 %, a stronger t-statistic of 5.08, and a superior Sharpe ratio of 0.20. Sectors like Utilities, Real Estate, Materials, and Financials, which experience lower volatility, display weaker performance, averaging annualised returns of 1.25%, a t-statistic of 1.40, and a Sharpe ratio of 0.07.

This means that while stocks may be more volatile, sentiment analysis is powerful at cutting through the noise created by high volatility. For practitioners, this suggests a strategic focus on sector volatility when selecting stocks for sentiment-based strategies.

5. Limitations and Future Research

5.1. Refinement of Four-Day Rolling Window

Our study utilised a four-day window to calculate cumulative abnormal returns (CARs), leading to insights such as the superior capability of sentiment analysis in identifying negative transcripts for short selling. This timeframe may skew results, as the market often reacts to bad news faster than good, with immediate stock price declines for negative sentiments observed within this window. In contrast, stock gains driven by positive sentiments might not be fully realised in the same period. Consequently, the pronounced effect on stock prices for negatively perceived companies is observed within this window, but the positive sentiment may yield stock appreciation that materialises beyond it, which may not be captured in our analysis. Future research should consider different timeframes for calculating returns: a shorter window may remain suitable for negative sentiment, while a longer window could better capture the eventual stock appreciation from positive sentiments.

Our study focused on dictionary development from the language used by management in the Q&A portions of earnings calls, often centred around financial health and company

performance as prompted by analysts' inquiries. For future studies, expanding the dictionary to include terms from the scripted parts of earnings calls could prove advantageous. These segments typically allow management to elaborate on industry trends and company performance, potentially offering a wealth of sector-specific terms. A deeper analysis of these scripted portions may yield a more targeted and detailed industry-specific lexicon, enhancing the dictionaries' specialisation and the effectiveness of sentiment analysis.

5.2. Segmenting Companies into Groups

Our study uses the Global Industry Classification Standard (GICS) to categorise companies into sectors for constructing industry-specific dictionaries. It is important to note that different methods to group companies may prove more effective due to some of the limitations of GICs. GICS classification is for 11 sectors, these are chosen based on rough approximations for where a company belongs. Company peer groups might be a better way to separate these companies. Standardised classification such as World Bank, may offer more granular groupings, potentially capturing linguistic patterns more representative of peer groups than GICS categories. This approach warrants exploration in future studies to enhance the accuracy of industry-specific sentiment analysis.

Furthermore, while some industry-specific dictionaries may overlap significantly with universal language sets, sectors like healthcare exhibit distinct terminologies unique to their field. A comparative analysis of industry languages to identify sectors with unique vocabularies could yield intriguing insights. Identifying which sectors have the most specialised language will assist in concentrating efforts on developing targeted dictionaries, thereby enhancing the precision and utility of sentiment analysis in those sectors.

5.3. Dictionaries Are Dynamic

Financial dictionaries are always evolving and can be characterised as 'dynamic'. This can be likened to a cat-and-mouse game where management shifts language towards a positive tone. To accurately reflect this evolving language, dictionaries must be updated continuously. Our study's approach—training the model on 80% of the data with a random seed—could introduce 'look ahead bias,' suggesting the need for ongoing updates to the dictionaries to maintain accuracy in sentiment definition. A potential improvement for future research is to segment the training and testing of models chronologically, ensuring that dictionaries are

trained on historical data before being tested on subsequent 'unseen' data. This time-based split could enhance the evaluation of a dictionary's predictive accuracy over time.

However, this approach is not without its challenges. Dictionaries calibrated on data up to 2021, for instance, may not accurately predict sentiment for 2022, due to the evolving nature of corporate communication. Future studies could investigate the optimal lifespan of a sentiment dictionary, determining how long it remains relevant before language shifts or new trends emerge, and a new dictionary has to be formed. Such research could reveal whether a dictionary's predictive validity extends to, say, a six-month forecast or if more frequent updates are required to align with management's change in language.

5.4. Dictionaries May Exhibit Seasonality

It is important to note that words and phrases in financial discourse may exhibit aspects of seasonality. Matsumoto et al. (2006) highlight the dynamic nature of earnings calls, identifying quarter-specific factors that influence the content and length of these calls. Future research can consider seasonality (quarter-specific factors). This approach can separate and update dictionaries on a same-quarter basis across different years, effectively tracking and capturing the evolving contexts and terminologies throughout quarters, pinpointing seasonal shifts in language patterns.

5.5. Sentiment vs Information

Our study focused on constructing a specialised lexicon derived from the qualitative content (Q&A sections from earnings calls) to assess sentiment in financial documents. A potential limitation, however, relates to the distinction between sentiment and tangible information within news content—a concept examined by von Beschwitz et al, (2015). They found that short sellers trade more on days with qualitative news due to its association with higher liquidity. Their findings suggest that on days with qualitative news, short selling is not necessarily information-driven but may exploit increased market liquidity.

In the context of our research, while we have developed dictionaries that can successfully measure sentiment, it is important to recognise that sentiment itself may not be a direct proxy for tangible information. The "news tangibility" concept from von Beschwitz et al, (2015) suggests that numerical content in news reports may offer clearer market expectations compared to the qualitative information used in this study.

For future research, it would be beneficial to investigate the relationship between the sentiment scores and the tangible, numerical information in financial texts. This research will help us understand whether sentiment scores, derived from qualitative information, represent 'noise trading' or carry 'tangible information'. By distinguishing the types of information that most impact market movements, subsequent studies could refine the predictive power of ML models for financial sentiment analysis.

6. Conclusion

In this study, we created industry-specific dictionaries using a Multinomial Naïve Bayes supervised machine learning approach and confirmed the presence of specialised language within sectors. We discovered that certain bigrams, such as “see disruption,” have different connotations across industries—being positive in Information Technology and negative in Consumer Discretionary. This finding demonstrates the limitations of a universal dictionary and supports the segmentation of dictionaries by industry for more accurate sentiment capture.

We provide insights on where ML-based sentiment analysis excels and how to use it effectively. Industries with higher volatility tend to provide better returns, suggesting that sentiment analysis tools are more effective in volatile environments. We show that in the short term, ML models excel at detecting negative sentiment and predicting stock price declines, providing an edge for practitioners who adopt a short-selling strategy.

While our research does not conclusively prove that industry-specific dictionaries are superior to universal dictionaries, it offers a strong proof-of-concept for segmenting lexicon-based sentiment analysis into industries. The exploration into industry-specific language lays the groundwork for both practitioners and future research, offering a new perspective on refining ML-based sentiment analysis.

Appendix

This table shows the top 50 positive and negative bigrams captured in the Universal Dictionary. The score represents the TF-IDF and Multinomial Naïve Bayes feature importance for each bigram, where higher sentiment scores would translate to a higher impact on stock prices.

Top 50 Positive and Negative Words

Negative		Positive	
Word	Score	Word	Score
cash burn	-0.516	good guidance	0.400
bottom end	-0.361	revenue management	0.345
license revenue	-0.327	demand strong	0.279
revenue decline	-0.326	incremental margin	0.260
combine company	-0.324	continue momentum	0.258
downward pressure	-0.320	really increase	0.253
obviously impact	-0.297	return invest	0.251
one expect	-0.296	momentum go	0.243
margin pressure	-0.287	growth margin	0.241
quarter total	-0.287	market leader	0.236
take inventory	-0.284	continuous improvement	0.235
expectation see	-0.273	expectation continue	0.233
maintenance capital	-0.261	get efficient	0.229
regulatory environment	-0.260	strong continue	0.225
product revenue	-0.257	tremendous growth	0.221
expect close	-0.251	pent demand	0.215
could also	-0.245	strong double	0.213

Negative		Positive	
next quarterly	-0.244	economic growth	0.211
think cover	-0.244	saw good	0.209
expect much	-0.236	continue execute	0.207
revenue business	-0.236	customer growth	0.201
cost pressure	-0.235	capital return	0.200
obviously also	-0.234	lot momentum	0.199
take longer	-0.230	well ahead	0.198
fuel cost	-0.228	see double	0.197
quarter work	-0.228	year high	0.197
impact say	-0.227	driver growth	0.196
acquisition cost	-0.222	growth give	0.196
actually sell	-0.221	margin opportunity	0.195
acquire business	-0.220	pay dividend	0.195
within guidance	-0.215	continue growth	0.194
negatively impact	-0.213	great start	0.192
access capital	-0.212	growth across	0.191
issue think	-0.210	invest capital	0.190
bad case	-0.205	strong cash	0.186
impact result	-0.203	company grow	0.185
expense line	-0.189	execute strategy	0.184
market improve	-0.186	great position	0.184
program also	-0.186	demand growth	0.182
remain optimistic	-0.185	high demand	0.179
meet expectation	-0.185	positive side	0.178
capital budget	-0.185	good quarter	0.178

Negative		Positive	
deferred revenue	-0.184	increase demand	0.177
sale product	-0.184	good execution	0.176
might take	-0.179	good cash	0.175
margin compression	-0.179	improve efficiency	0.174
impact revenue	-0.177	strong performance	0.174
think difficult	-0.176	grow earnings	0.174
low gross	-0.175	maintain margin	0.171
labor cost	-0.173	strong revenue	0.169

This analysis was based on if one word in the bigram matches with the universal dictionary. This shows where industry dictionaries share common bigrams with the universal dictionary, and also how specialised certain industries are compared to the universal dictionaries.

Industry-Specific vs Universal Dictionary

Industry	Common Bigrams with Universal	Unique Bigrams
Information Technology	64%	36%
Consumer Discretionary	65%	35%
Health Care	57%	43%
Industrials	69%	31%
Financials	56%	44%
Energy	50%	50%
Materials	56%	44%
Real Estate	49%	51%
Consumer Staples	55%	45%
Communication Services	50%	50%
Utilities	38%	62%

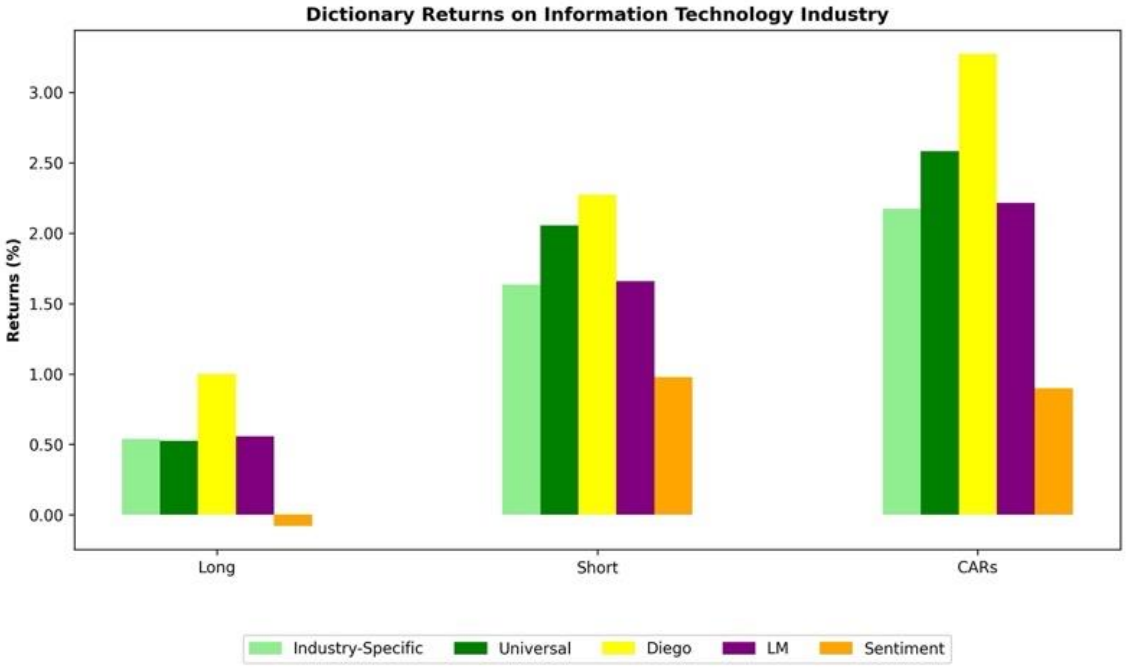
Our analysis of industry-specific dictionaries, particularly in the Information Technology (IT), Consumer Discretionary (CD), and Health Care (HC) sectors, yields insightful findings regarding the occurrence of common and unique bigrams compared to our universal dictionary. For instance, in the IT sector, we discovered that 64% of the bigrams are common with the universal dictionary, while 36% are unique to IT. This distribution of bigram commonality offers an interesting perspective on the degree of specialized language used in different industries. For comparison, the CD sector exhibits 35% unique bigrams relative to the universal dictionary, whereas the HC sector shows a higher proportion, with 43% unique bigrams. These percentages suggest that the HC sector may employ more specialized language. This preliminary observation opens avenues for further research to delve deeper into the distinct linguistic features of each industry and their implications for sentiment analysis.

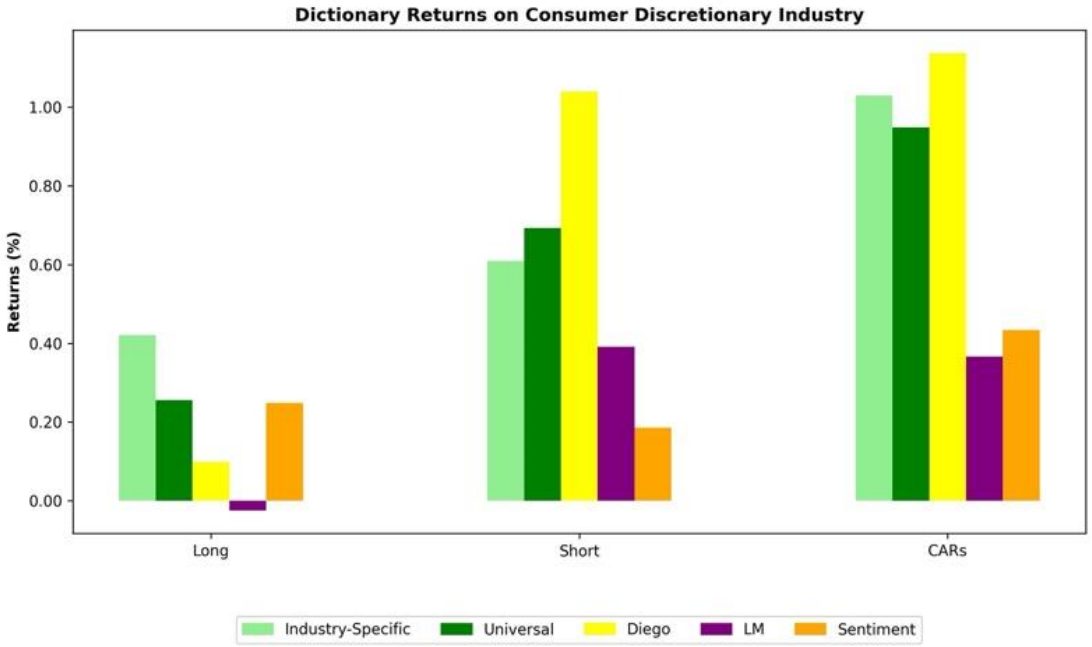
This table describes the Long Returns, Short Returns, and CARs (L-S Returns) for all dictionaries across each industry. The table also shows each sector's volatility from out-of-sample companies in that specific industry. Volatility is calculated as the standard deviation of CARs for all the out-of-sample companies in the corresponding industry.

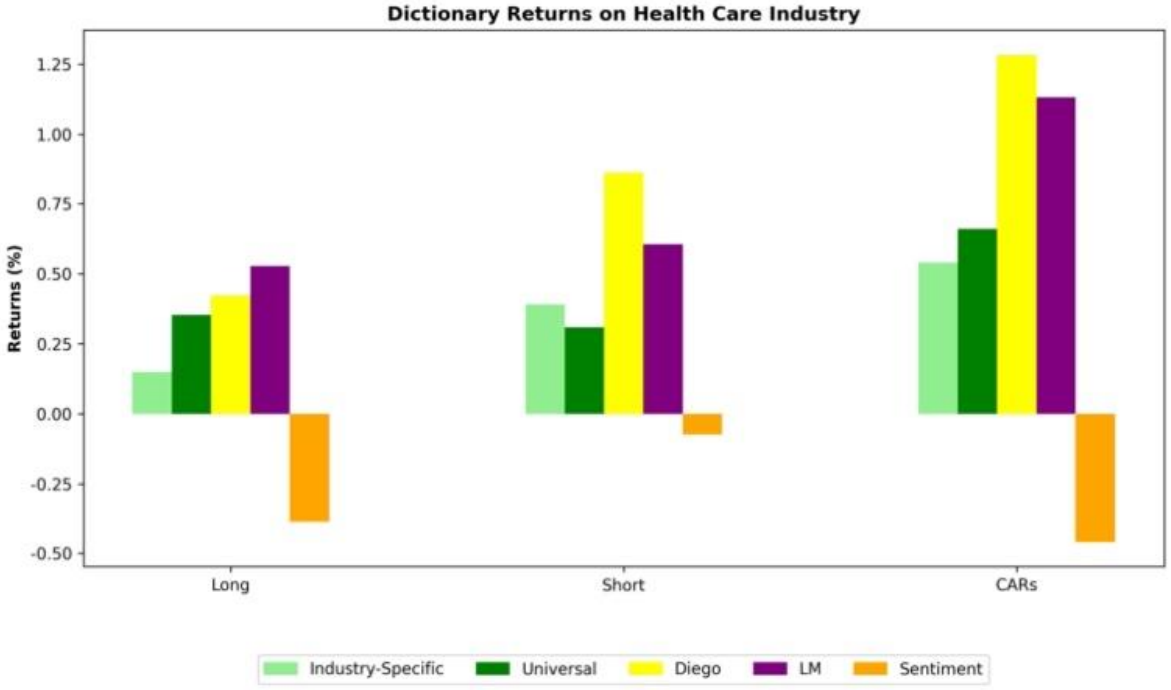
Sector	Industry-specific	Universal (0.25)	Diego	LM	Sentiment	Volatility
	LR SR CARs (t-stat)					
IT	0.54% -1.62% 2.16% (8.41)	0.53% -2.05% 2.58% (10.12)	1.00% -2.27% 3.27% (12.80)	0.56% -1.66% 2.21% (8.53)	-0.08% -0.98% 0.90% (3.39)	7.72
Consumer Discretionary	0.42% -0.61% 1.03% (4.80)	0.26% -0.69% 0.95% (4.34)	0.10% -1.04% 1.14% (5.38)	-0.02% -0.39% 0.37% (1.63)	0.25% -0.19% 0.43% (2.00)	6.25
Health Care	0.15% -0.39% 0.54% (2.04)	0.35% -0.31% 0.66% (2.62)	0.42% -0.86% 1.29% (5.22)	0.53% -0.61% 1.13% (4.19)	-0.39% 0.07% -0.46% (-1.71)	7.09
Industrials	-0.27% -0.38% 0.10% (0.51)	0.13% -1.01% 1.14% (5.54)	0.23% -1.00% 1.24% (6.11)	-0.22% -0.19% (0.03%) (-0.16)	-0.16% -0.56% 0.40% (1.98)	5.58
Financials	0.28% 0.22% 0.05% (0.31)	0.47% -0.28% 0.75% (4.48)	0.67% -0.34% 1.01% (5.76)	0.29% 0.28% 0.00% (0.01)	0.30% -0.19% 0.49% (2.85)	4.38
Energy	-0.62% -1.58% 0.96% (2.95)	-0.58% -1.76% 1.18% (3.79)	-0.88% -1.14% 0.27% (0.81)	-0.13% -0.73% 0.60% (1.96)	-0.99% -0.48% (0.51%) (-1.69)	5.76
Materials	0.42% -0.10% 0.52% (1.78)	-0.05% 0.49% (0.54%) (-1.92)	0.01% -0.44% 0.45% (1.52)	-0.14% -0.58% 0.45% (1.61)	0.04% -0.60% 0.64% (2.28)	4.32

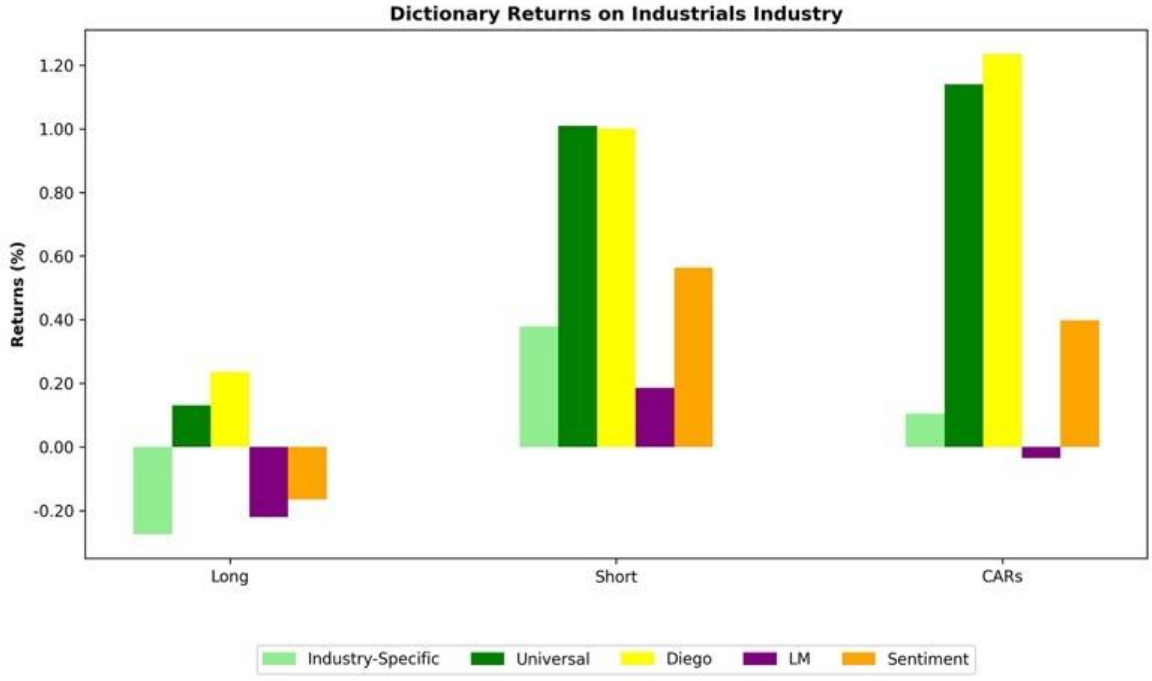
Sector	Industry-specific	Universal (0.25)	Diego	LM	Sentiment	Volatility
	LR SR CARs (t-stat)					
Real Estate	-0.41% 0.09% -0.49% (-2.30)	0.38% -0.74% 1.12% (4.41)	0.46% -0.81% 1.27% (4.99)	0.14% -0.28% 0.42% (1.61)	-0.22% -0.56% 0.34% (1.31)	3.95
Consumer Staples	-0.09% 0.03% (0.11%) (-0.33)	0.18% -0.89% 1.07% (2.96)	0.78% -0.73% 1.51% (3.92)	0.32% -1.43% 1.74% (5.08)	-0.56% -0.77% 0.21% (0.62)	5.42
Communication Services	0.93% -1.16% 2.10% (4.76)	0.92% -0.36% 1.28% (2.69)	0.35% -1.67% 2.01% (4.34)	-0.83% -0.57% (0.26%) (-0.56)	-0.35% -0.32% (0.03%) (-0.06)	6.59
Utilities	-0.22% -0.04% (0.17%) (-0.49)	-0.13% 0.87% (1.00%) (-3.29)	-0.12% -0.58% 0.46% (1.37)	0.91% 0.24% 0.68% (2.29)	0.37% 0.08% 0.29% (1.07)	3.45

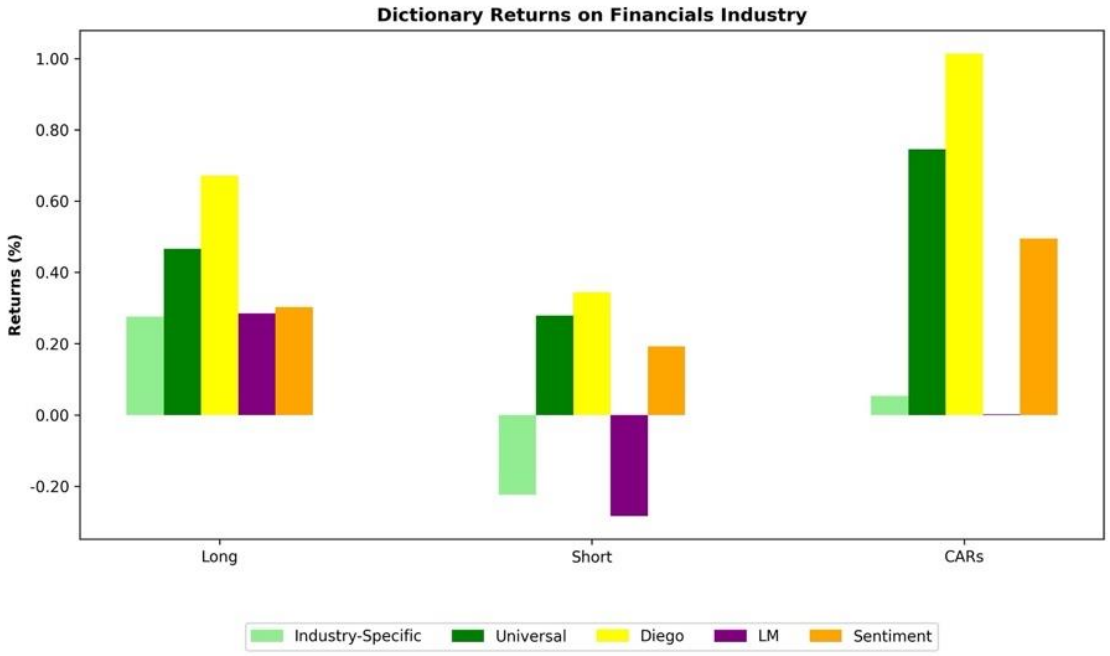
These figures represent the Long Returns, Short Returns, CARs (L-S Returns) for all the dictionaries across each industry.

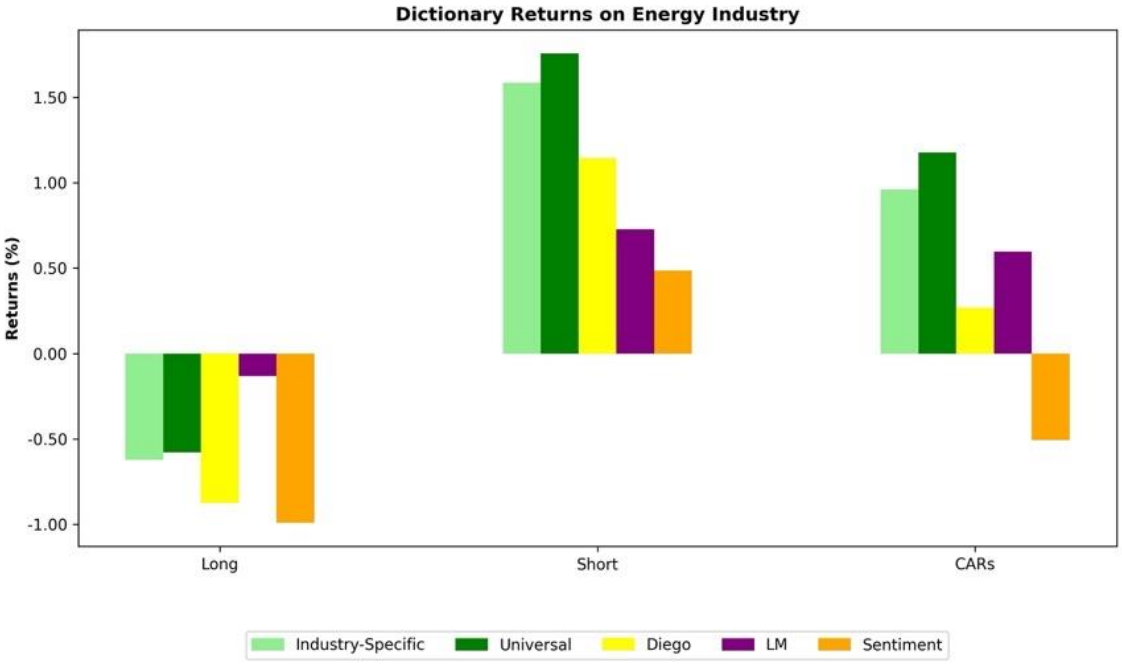


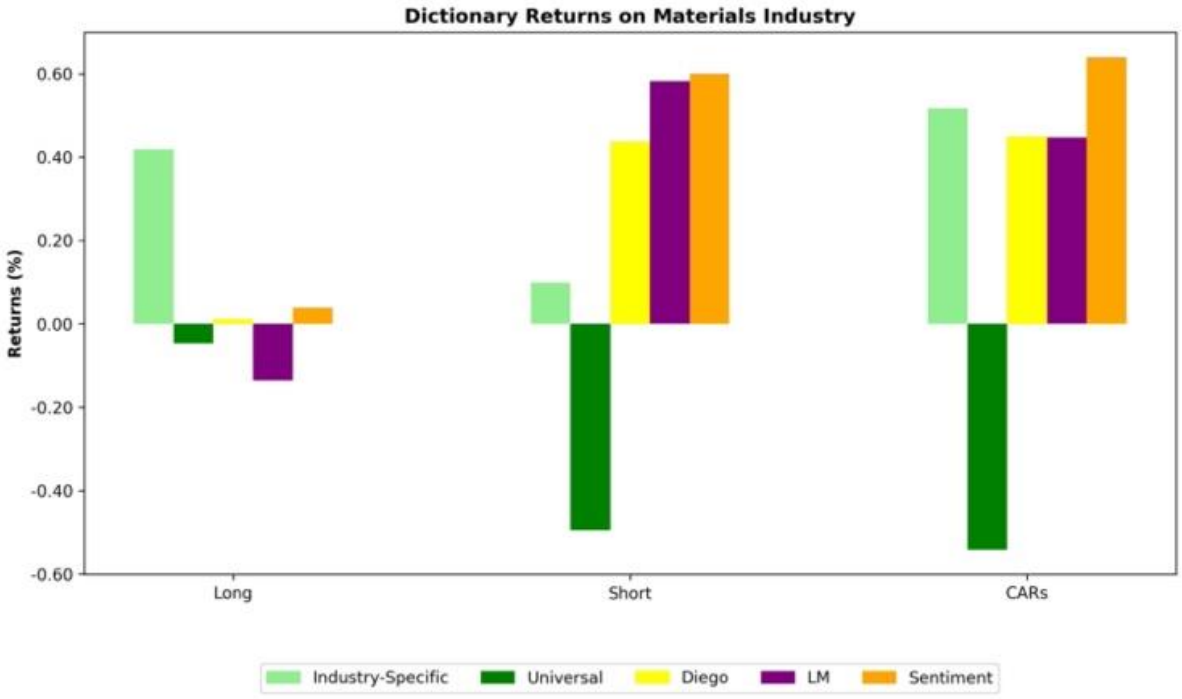


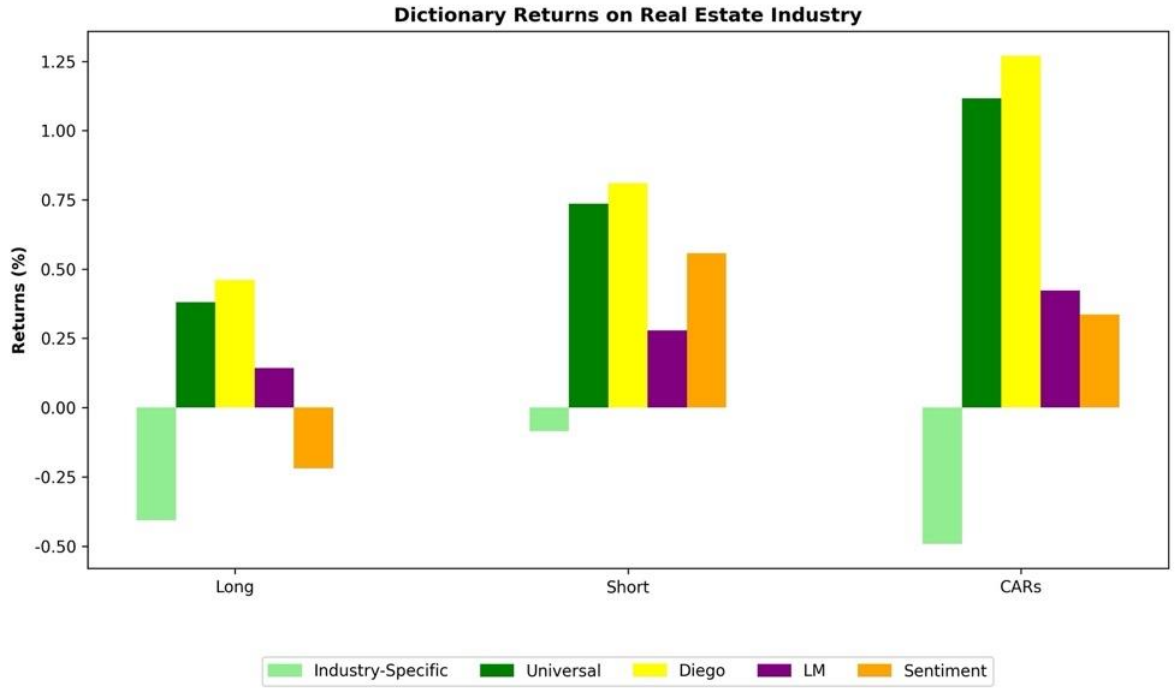

Dictionary Returns on Health Care Industry
Returns (%)
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment

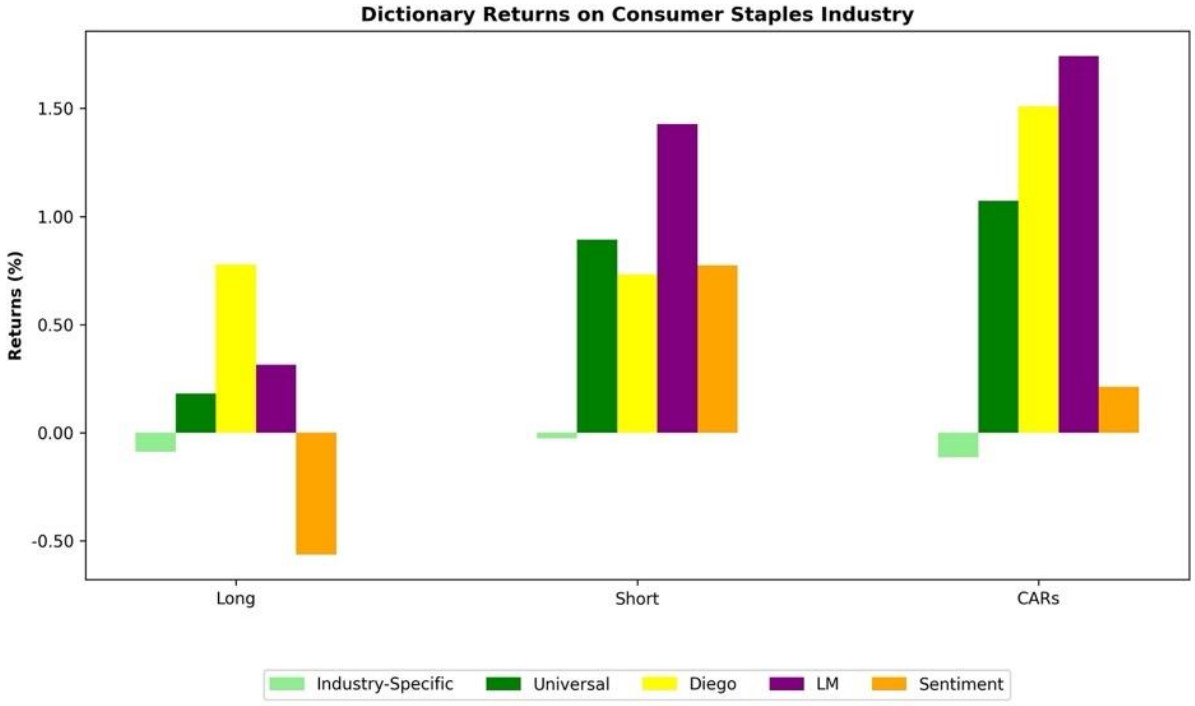

Dictionary Returns on Industrials Industry
Returns (%)
1.20
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment

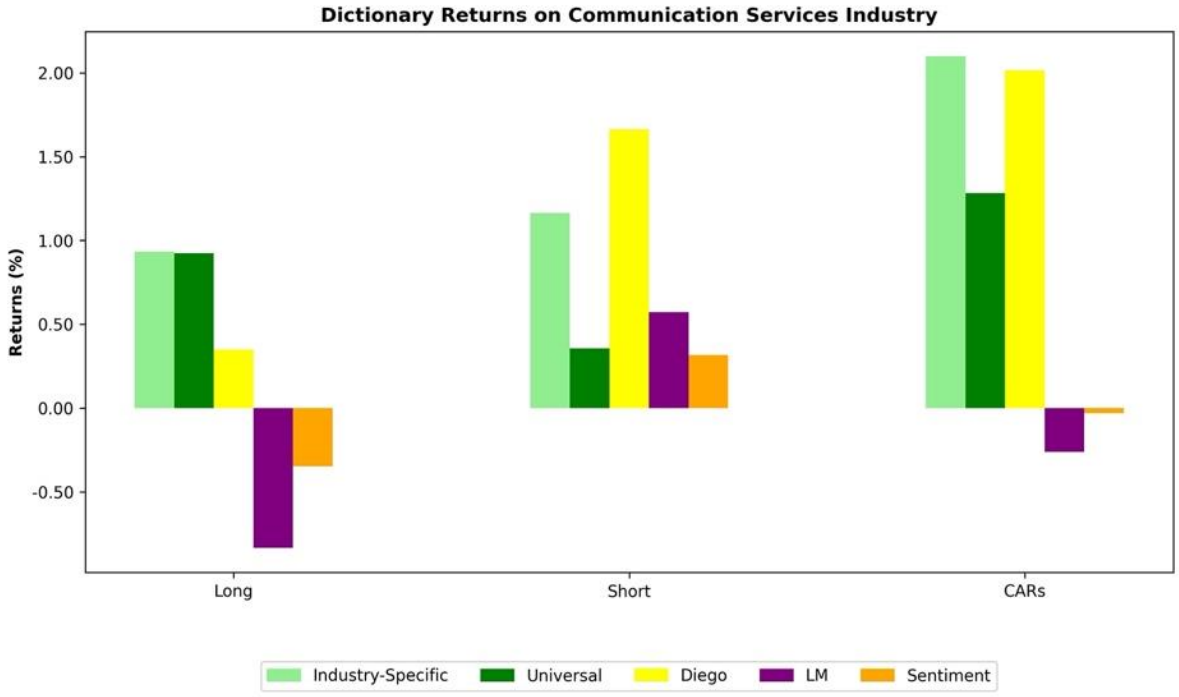

Dictionary Returns on Financials Industry
Returns (%)
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment

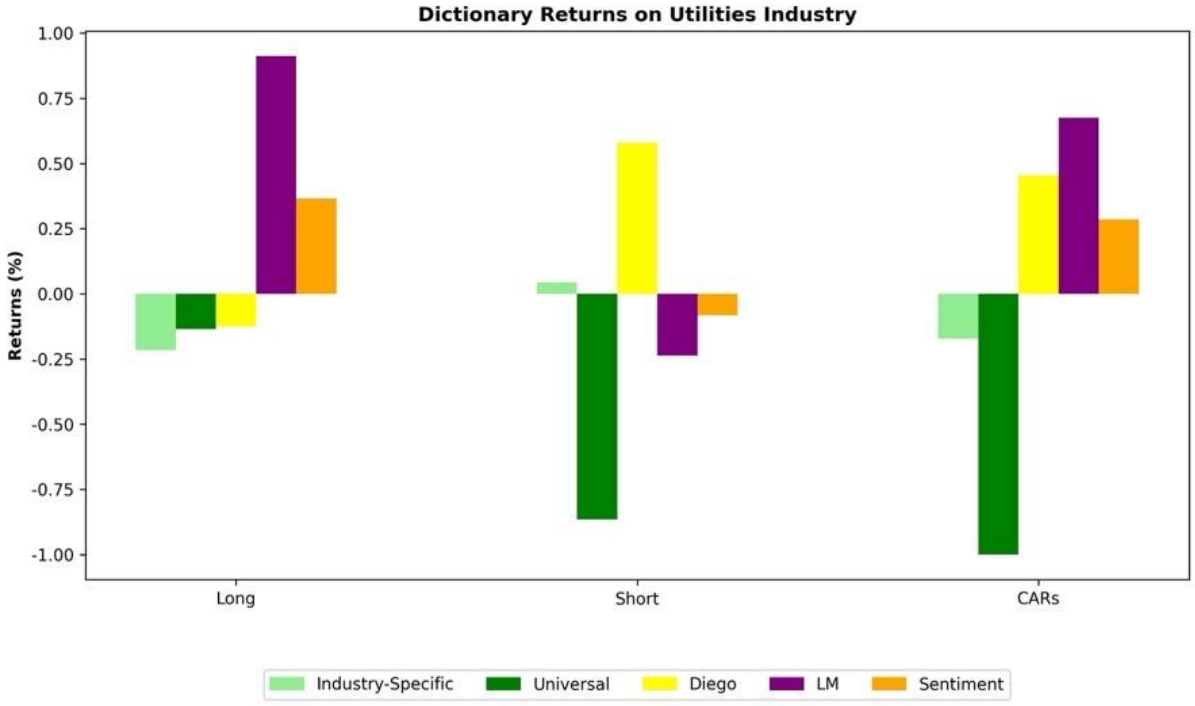

Dictionary Returns on Energy Industry
Returns (%)
1.50
1.00
0.50
0.00
-0.50
-1.00
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment


Dictionary Returns on Materials Industry
Returns (%)
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment


Dictionary Returns on Real Estate Industry
Returns (%)
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment


Dictionary Returns on Consumer Staples Industry
Returns (%)
1.50
1.00
0.50
0.00
-0.50
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment


Dictionary Returns on Communication Services Industry
Returns (%)
2.00
1.50
1.00
0.50
0.00
-0.50
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment


Dictionary Returns on Utilities Industry
Returns (%)
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
Long
Short
CARs
Industry-Specific
Universal
Diego
LM
Sentiment

This figure shows the returns for dictionaries CARs across different proportions of out-of-sample data. The percentage of out-of-sample data equal to 2% represents 240 companies in a portfolio; it only encapsulates the most positive and negative transcripts, whereas 100% captures the whole out-of-sample data.

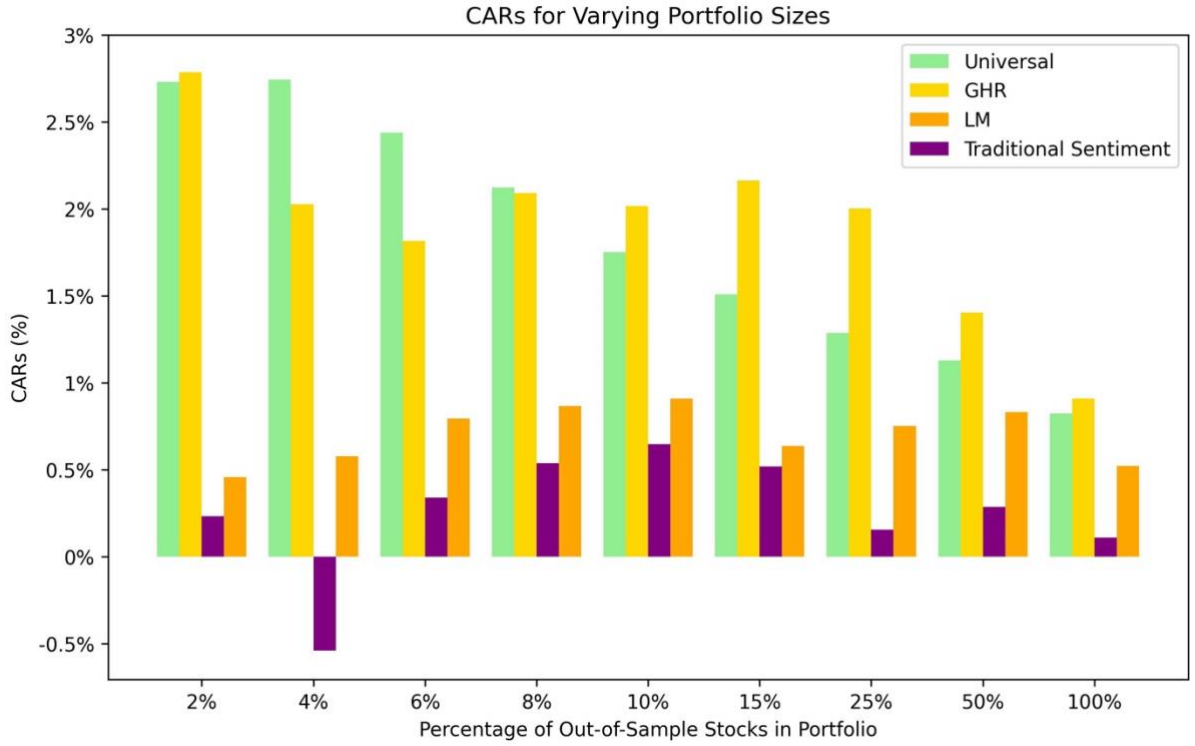


Figure 4: Dictionary Performance Across Portfolio Sizes

References

Antweiler, W., & Frank, M. Z. (2006). Do us stock markets typically overreact to corporate news stories? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.878091

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1908.10063

Bachhety, S., Dhingra, S., Jain, R., & Jain, N. (2018). Improved Multinomial Naïve Bayes Approach for Sentiment Analysis on Social Media. SSRN. https://ssrn.com/abstract=3363601

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K Text to Gauge Financial Constraints. Journal of Financial and Quantitative Analysis, 50(4), 623–646. https://doi.org/10.1017/S0022109015000411

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. The Review of Financial Studies, 36(9), 3603–3642. https://doi.org/10.1093/rfs/hhad021

Consoli, S., Barbaglia, L., & Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. Knowledge-Based Systems, 247, 108781-. https://doi.org/10.1016/j.knosys.2022.108781

García, D., Hu, X., & Rohrer, M. (2023). The colour of finance words. Journal of Financial Economics, 147(3), 525–549. https://doi.org/10.1016/j.jfineco.2022.11.006

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

Griffin, P. A. (2003). Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings. Review of Accounting Studies, 8(4), 433-. https://doi.org/10.1023/A:1027351630866

Heston, S. L., & Sinha, N. R. (2017). News vs. Sentiment: Predicting Stock Returns from News Stories. Financial Analysts Journal, 73(3), 67–83. https://doi.org/10.2469/faj.v73.n3.3

Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. The Journal of Political Economy, 124(5), 1423–1465. https://doi.org/10.1086/688176

Hu, W., Shohfi, T., & Wang, R. (2021). What's really in a deal? Evidence from textual analysis of M&A conference calls. Review of Financial Economics, 39(4), 500–521. https://doi.org/10.1002/rfe.1126

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

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

Ke, Z. T., Kelly, B., & Xiu, D. (2019). Predicting Returns with Text Data. https://doi.org/10.3386/w26186

Kelly, B., Manela, A., & Moreira, A. (2021). Text Selection. Journal of Business & Economic Statistics, 39(4), 859–879. https://doi.org/10.1080/07350015.2021.1947843

Liang, P. J., Meursault, V., Routledge, B. B., & Madeline Marco Scanlon. (2021). PEAD.txt: Post-Earnings-Announcement Drift Using Text. IDEAS Working Paper Series from RePEc. https://doi.org/10.21799/frbp.wp.2021.07

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.07619

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

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Loughran, T., & McDonald, B. (2020). Measuring firm complexity. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3645372

Loughran, T., & McDonald, B. (2020). Textual Analysis in Finance. Annual Review of Financial Economics, 12(1), 357–375. https://doi.org/10.1146/annurev-financial-012820-032249

Matsumoto, D. A., Roelofsen, E., & Pronk, M. (2006). Managerial disclosure vs. analyst inquiry: An empirical investigation of the presentation and discussion portions of earnings-related conference calls. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.943928

Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion

Sessions. The Accounting Review, 86(4), 1383–1414. https://doi.org/10.2308/accr-10034

Narayana, D., Anwesh, R.P., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U., & Agarwal, Y. (2022). Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2204.05783

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

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

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. Journal of the American Statistical Association, 108(503), 755–770. https://doi.org/10.1080/01621459.2012.734168

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance (New York), 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Sarr-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance (New York), 63(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

von Beschwitz, B., Chuprinin, O., & Massa, M. (2017). Why Do Short Sellers Like Qualitative News? Journal of Financial and Quantitative Analysis, 52(2), 645–675. https://doi.org/10.1017/S0022109017000151

Zaremba, A., & Demir, E. (2023). CHATGPT: Unlocking the future of NLP in Finance. Modern Finance, 1(1), 93–98. https://doi.org/10.61351/mf.v1i1.43