HEC MONTRÉAL

Strategies to Monetize the Sentiment Extracted With NLP Techniques From Earnings Call Transcripts

by

Benjamin Séguin

Under the supervision of

David Ardia

Department of Decision Sciences
(Financial Engineering)

In Partial fulfillment of the requirements for
the Degree of Master of Science (M.Sc.)

April 2025

Abstract

This study explores portfolio construction by leveraging sentiment analysis from earnings call transcripts of North American companies in the software sector. Using advanced Natural Language Processing (NLP) techniques and a basic lexicon-based approach with the Loughran and McDonald (2011) dictionary, I analyze sentiment at various levels of granularity. Sentiment is computed using a Natural Language Inference (NLI) approach with DeBERTa (He et al., 2021) and a straightforward method with FinBERT (Araci, 2019). Initially, I segment the transcripts into multiple parts, create summaries, and extract sentiment for each segment before aggregating them into a final score to build diverse portfolio strategies. Subsequently, I combine these individual summaries to generate a comprehensive sentiment score for the entire transcript, providing a broader sentiment assessment. My research includes long-short and long-only frameworks, different weighting schemes, and several sentiment scores aggregation methodologies. I find that the granularity of sentiment extraction significantly impacts portfolio performance, and I demonstrate the superiority of this approach in a long-only framework. I propose several aggregation methods for sentiment scores, and I find that the naive-average methodology is the most effective. Finally, I show that a sentiment-weighted portfolio construction yields better results than a classic equal-weight approach.

JEL Classification: C11; C12; C22; C32; C52; C53.

Keywords: Natural Language Processing; Natural Language Inference; Large Language Models; Sentiment Analysis; Portfolio Construction

Acknowledgments

First and foremost, I would like to express my sincere gratitude to the Caisse de Dépôt et Placement du Québec (CDPQ) for the opportunity to collaborate with them on this project, utilizing their data and building upon some of their foundational work on the coding side. In particular, I extend my heartfelt thanks to Marie-Eve Malette and Vincent Gariépy for their invaluable guidance and support throughout the course of this research.

I am also deeply thankful to my friends and family for their unwavering support, for listening to me explain my work—even when it was neither easily understood nor of particular interest to them—and for always being there for me.

Finally, I would like to extend my deepest appreciation to my supervisor and mentor, David Ardia, for his generous investment of time and insightful guidance. The path to completing this master's thesis has been filled with uncertainty, as I changed my research topic multiple times before settling on this final iteration. I am immensely grateful for his patience, encouragement, and for helping me navigate the fascinating world of academic research.

1 Introduction

As finance continues to evolve into a data-driven discipline, we are rapidly shifting from traditional structured datasets to the vast and complex realm of unstructured data. The importance of this shift is widely discussed in the literature (see Cong et al., 2019, Gentzkow et al., 2019, Loughran and McDonald, 2020), and has become even more pronounced with the emergence of Large Language Models (LLMs). A general conclusion of these studies is that the use of advanced techniques like Natural Langugage Processing (NLP) on unstructured data allow to extract new usable information for investment purposes.

The Efficient Market Hypothesis (EMH) posits that financial markets are efficient in reflecting all available information in the prices of securities (Fama, 1970). However, subsequent research, such as Grossman and Stiglitz (1980), challenges this notion by arguing that markets may not be perfectly efficient. They suggest that some publicly available data may contain latent information not immediately evident to all market participants. This perspective opens the door to exploring various market inefficiencies, particularly the potential for certain types of information like textual content, to be underutilized by investors. Li (2006) demonstrate that textual information is often less incorporated into market valuations, pointing to a possible oversight in the pricing of securities.

Against this backdrop, textual sentiment analysis emerges as a powerful tool to understand and potentially exploit the psychological foundations of investor behavior. By quantifying the affective content of textual data, sentiment analysis aims to capture the cognitive biases that influence market prices (Tetlock, 2007).

Sentiment analysis has thus become a focal point within the financial sector, offering an informational edge. Early work by Hu and Liu (2004) establish a foundation for extracting sentiment from textual data, which has been adapted for financial documents to generate market insights (Tetlock et al., 2008). Subsequent research, such as Kogan et al. (2009), expand on these methods by employing regression techniques to assess risk from the language in corporate financial reports, demonstrating the predictive power of textual sentiment analysis in financial contexts. Bollen et al. (2011) further illustrate how aggregated mood data from social media could predict stock market trends, suggesting that sentiment analysis can provide early indicators of market movements.

The literature prior to 2013 largely relied on regression techniques and basic machine learning algorithms for sentiment analysis, as summarized in Kearney and Liu (2014). However, the advent of deep learning has significantly advanced the field. Kim (2014) demonstrates the application of Convolutional Neural Networks (CNNs) to sentiment analysis tasks, while Tang et al. (2015) explore the use of Long Short-Term Memory (LSTM) networks and Gated Recurrent Neural Networks (GRNNs) for more accurate sentiment analysis. Chakraborty et al. (2019) further combine CNNs and LSTMs in a hybrid model, showing that deep learning models generally outperform traditional machine learning algorithms and dictionary-based approaches due to their ability to capture semantic nuances within texts.

The development of transformer architectures has marked a revolutionary leap forward in NLP. Introduced by Vaswani et al. (2017), transformers use self-attention mechanisms to weigh the importance of different words within a sentence, leading to a deeper understanding of context and semantics. This innovation paved the way for LLMs such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), which have set new benchmarks across a wide range of NLP tasks. The ability of transformers to process and generate human-like text has significantly enhanced sentiment analysis, allowing for more nuanced interpretations of textual data.

More recent studies have leveraged these advancements for various predictive tasks in finance. For example, Jha et al. (2024) use ChatGPT to generate firm-level scores from conference call transcripts to forecast capital expenditure adjustments, while Lopez-Lira and Tang (2023) and Pelster and Val (2024) demonstrate the predictive power of ChatGPT-4 in analyzing news headline sentiments and aiding stock selection. Fatouros et al. (2023) show the effectiveness of ChatGPT-3.5 in financial sentiment analysis, and Schuettler et al. (2024) fine-tune their own LLM to perform sentiment analysis and create long/short portfolios, both highlighting the growing role of LLMs in finance. Lopez-Lira and Tang (2023) also demonstrate the capability of LLMs to forecast stock prices, classifying stocks as long, uncertain, or short based on news headlines, and then using these predictions to forecast stock returns. Lefort et al. (2024) use Bloomberg news headlines, as a data source. They initially filter it to retain only those with potential market impact and they ask ChatGPT to classify each news item as likely to cause an increase, decrease, or have a neutral effect on financial markets. Finally, Chen et al. (2022) derive sentiment from Reuters news for individual stocks and find that using LLMs to construct sentiment-based portfolios outperforms any other sentiment analysis technique.

I build on these recent advancements in NLP and LLMs to analyze earnings call transcripts and introduce a new methodology. My approach leverages GPT-3.5 for its text summarization capabilities, as demonstrated by Brown et al. (2020) and Kim et al. (2024), and integrates advanced prompt engineering techniques from Yue et al. (2023) and Bommarito and Katz (2022). I further enhance sentiment extraction using DeBERTa, which improves the understanding of word relationships (He et al., 2021), and apply NLI for nuanced interpretation. Drawing on the sentiment-based portfolio strategies of Wang et al. (2018), which show the effectiveness of portfolio construction based on sentiment, my research seeks to profit from this extracted information.

In my novel methodology, I initially divide the earnings call transcripts into chunks of maximum 8000 tokens each. I choose this number because this is the maxmimum size that can be processed by GPT-3.5 at once. After segmentation, I summarize each of these chunks using GPT-3.5; I will refer to these summaries as 'atoms.' Then, I compute sentiment scores using three distinct methodologies: DeBERTa (He et al., 2021) in a NLI framework, FinBERT (Araci, 2019), which provides discrete sentiment scores, and a more traditional lexicon-based approach using the Loughran and McDonald (2011) dictionary. Additionally, I explore varying levels of granularity. In the first one, I compute sentiment scores for each atom and explore different methods to aggregate the atoms' scores into a final score for the transcript. In the second one, I compute a single score on a single summary of the whole transcript.

Then, I monetize the information extracted by constructing both long-only and long-short portfolios that are rebalanced each quarter. The long-only portfolios are constructed by buying the 10 stocks with the highest sentiment score. The long-short portfolios have the same long portion, but they are also made of a short position in the 10 stocks with the lowest sentiment scores. I also explore different methodologies to weight the stocks in the portfolios, and I evaluate each combination of methods with traditional portfolio performance metrics such as the Sharpe ratio and the alpha.

For the empirical analysis, I utilize a dataset of 69 companies from the software sector, covering a backtesting period from Q1 2013 to Q4 2023, based on calendar quarters. I have access to the complete history of earnings call transcripts from 2013 onward, as well as monthly stock prices, which I convert into quarterly returns for each company. Some companies in my dataset either went public after 2013 or were delisted due to privatization or bankruptcy before the end of 2023; these adjustments have been fully accounted for and will be described in greater detail later in the analysis.

The use of LLMs and advanced NLP techniques in sentiment analysis represents a rapidly evolving field with significant potential for future research. My study contributes to this growing body of literature by introducing a continuous sentiment scoring mechanism via NLI, which I compare against traditional binary classification methods, such as those provided by Loughran and McDonald (2011) and Araci (2019). I show that my model outperforms the binary ones in a long-only setting in terms of absolute and risk-adjusted performance metrics. With my best strategies, I am able to achieve a cumulative gain of 18.35$ and Sharpe ratio of 2.36 with DeBERTa, whereas I get 17.29$ and 1.34 for FinBERT and 13.71$ and 1.68 for the dictionary method. However, DeBERTa strategies still lag behind FinBERT in the long-short framework. I also find that in terms of portfolio construction, weighting the stocks by their respective sentiment scores outperforms the traditional equal-weight framework.

My research demonstrates the advantages of the atom methodology in capturing incremental information from textual data. This innovative approach lays the groundwork for future investigations into more granular sentiment analysis techniques. By exploring various sentiment score aggregation strategies, I provide a comprehensive framework for applying advanced NLP tools in financial analysis, opening new avenues for both academic research and practical application in quantitative finance.

My key takeaways are that using an NLI framework to construct a continuous sentiment score increases the precision and thus is better suited to use in a financial context for signal generation. Used this way, I can also conclude that DeBERTa provides better buy signals than FinBERT even though this latter LLM was trained on a corpus of financial texts. Another takeaway is that the granularity at which I compute the score matters and, in my case, the more granular strategies yield better signals. Finally, the sentiment-weighting methodology should be applied when constructing portfolios with sentiment scores, as it always outperforms the equal-weight methodology.

This thesis is organized as follows. Section 2 presents the data. Section 3 describes the NLP methodology. Section 4 presents the portfolio construction. Section 5 shows the empirical results. Finally, Section 6 concludes and presents the limitations of my research and possible extensions for future research.

2 Data

This section provides an overview of the data utilized in my analysis. The textual data comprises transcripts of earnings calls from software companies, which are employed to compute sentiment scores. For portfolio construction, I extract price data from Bloomberg Terminal, while factor data and the risk-free rate are sourced from the Kenneth R. French Data Library.

2.1 Earnings Call Transcripts

Data collection begins with the extraction and processing of transcripts from earnings calls of 69 North American companies in the software sector (GICS code: 451030), covering the period from January 1, 2013, to December 31, 2023. In total, 2,068 distinct transcripts are processed. The investment universe is defined to align with the transcript universe available in the CDPQ database. It should be noted that not all companies in the dataset are continuously present or publicly traded throughout the entire study period, resulting in fewer than 44 transcripts—the total number of quarters analyzed—for certain firms. Due to confidentiality agreements and data ownership by CDPQ, further details regarding the distribution of transcripts over time or by firm cannot be disclosed. The resulting imbalance in transcript availability does not compromise the validity of the analysis, as it is systematically addressed during stock selection and benchmark construction.

To process these transcripts effectively, I segment them into chunks of approximately 1000 tokens each. Although the model used, GPT-3.5 by OpenAI (Brown et al., 2020) can receive up to 8000 tokens, I decide to use 1000 tokens only to have a more granular approach and have more details in the summary. I use GPT-3.5 due to its cost-effectiveness compared to more advanced models and its availability through the CDPQ. In the domain of NLP, the concept of a 'token' is central to many computational linguistics tasks. A token can be broadly defined as a meaningful unit of text, typically a word, part of a word, or a punctuation mark, that serves as input for further processing as discussed in Grefenstette (1999).

After segmenting the text into chunks, I generate summaries for each chunk, referred to as 'atoms' and denoted by A_i. These atoms are constructed using GPT-3.5 in a direct manner, requiring no preprocessing steps such as stemming, lemmatization, or token reduction. Each chunk is input to the LLM with a prompt to generate a concise summary of fewer than 100 words. The original transcripts, which range in length from 5,000 to approximately 16,000 tokens, are thus decomposed into sets of atoms. For a given transcript T_i the resulting set of atoms is denoted as $A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}\}$ where n_i (the number of atoms per transcript) varies between 7 and 15, depending on transcript length. Aggregating across all transcripts in the dataset, the total number of atoms generated for all companies and time periods is 22,976. Due to confidentiality agreements, I cannot say more about the proprietary data used in this analysis.

2.2 Prices and Factors

I collect the closing prices of the 69 stocks at the end of each quarter to calculate simple quarterly returns. This approach ensures consistency with the timing of portfolio return calculations, which are also based on quarter-end values. As data availability for each stock varies over time, the breadth of the investment universe fluctuates across different periods. Nevertheless, this variation does not adversely affect my portfolio construction methodology. Both benchmarks and portfolios are constructed using all available stocks at each point in time. I also collect quarterly risk-free rate and factor data to compute performance metrics, including the Sharpe ratio and the alpha derived from my portfolio's regression on the FF six-factor model.

3 NLP Methodology

In this section, I introduce the methodology underlying LLMs. I provide an overview of transformer architectures, explain the differences between BERT, FinBERT, and DeBERTa, and introduce NLI.

3.1 Transformer Architecture and Core Principles

Transformers (Vaswani et al., 2017) revolutionized neural network design by replacing sequential processing with parallelized attention mechanisms. Building on this foundation, modern LLMs like BERT and DeBERTa implement specialized adaptations for financial text analysis.

3.1.1 Core Components

The transformer architecture consists of three primary components:

- **Token Embeddings**: Convert discrete text tokens into continuous vector representations using learned embeddings ($E \in \mathbb{R}^{d_{\text{model}}}$), where d_{model} typically ranges from 512 to 4096 dimensions in modern implementations.

- **Positional Encodings**: Inject sequence order information through sinusoidal functions:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \tag{1}$$

  where pos is the position index and i is the dimension.

- **Multi-Head Attention**: Enables simultaneous focus on different contextual relationships through h parallel attention heads ($h = 8$ in original implementation).

3.1.2 Self-Attention Mechanism

The scaled dot-product attention computes contextual relationships between all token pairs:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{'}}{\sqrt{d_k}}\right)V. \tag{2}$$

Where:

- Q (Query): Represents the current token's embedding seeking context.

- K (Key): Represents embeddings of all tokens in the sequence.

- V (Value): Represents content associated with each token.

- $\sqrt{d_k}$: A scaling factor that prevents excessively large gradients when d_k (dimensionality of key vectors) is high.

This mechanism allows transformers to dynamically focus on relevant words in a sequence. For example, in an earnings call transcript, the model can prioritize phrases like “margin expansion” while downplaying less relevant boilerplate text.

3.1.3 Advantages Over Sequential Models

Transformers offer several advantages over earlier architectures like RNNs and LSTMs:

- **Parallelization**: Processes all tokens simultaneously, significantly improving computational efficiency.
- **Long-Range Context**: Captures dependencies across entire sequences, enabling nuanced understanding of complex texts.
- **Transfer Learning**: Pre-training objectives such as Masked Language Modeling (MLM) allow transformers to generalize across diverse NLP tasks.

Transformers have proven particularly effective in financial NLP tasks such as semantic parsing, event extraction, and cross-document reasoning.

For a more detailed mathematical treatment of transformers, readers can refer to the work of Phuong and Hutter (2022).

3.2 Bert: Bidirectional Encoder Representations From Transformers

While transformers revolutionized sequence processing, BERT (Devlin et al., 2019) introduced bidirectional context understanding through two key innovations:

3.2.1 Core Architecture

The BERT architecture consists of:

- **Token Embeddings**: Maps input tokens to d_{model}-dimensional vectors ($d_{\text{model}} = 768$ in base BERT).
- **Positional Encodings**: Injects sequence order information through learned positional embeddings.
- **Transformer Layers**: Stacked self-attention and feed-forward networks (12 layers in base BERT).

3.2.2 Training Objectives

BERT's pre-training uses two unsupervised tasks to learn contextual relationships in text: MLM and Next Sentence Prediction (NSP).

MLM MLM trains BERT to recover masked tokens by analyzing bidirectional context. For input sequence $x = [x_1, ..., x_n]$, the model randomly masks 15% of tokens. The masking strategy is as follows:

- 80% of masked tokens are replaced with the [MASK] token.
- 10% are replaced with random tokens.
- 10% remain unchanged.

This ensures that the model does not overfit to the [MASK] token during fine-tuning.

The MLM objective function is defined as:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{x \sim \mathcal{D}} \sum_{i \in M} \log P(x_i | x_{\setminus M}), \tag{3}$$

where M is the set of masked positions, $x_{\setminus M}$ is the sequence with masked tokens, and $\mathcal{D}$ is the training corpus.

For example:

- Original sentence: "Revenue grew 15% despite macroeconomic headwinds."
- Masked sentence: "Revenue [MASK] 15% despite [MASK] headwinds."
- The model learns to predict 'grew' and 'macroeconomic' using surrounding context.

MLM enables BERT to understand semantic relationships between words, making it particularly useful for financial contexts where nuanced language often conveys critical information.

NSP NSP trains BERT to understand inter-sentence relationships, which is crucial for tasks like analyzing earnings call transcripts where context spans multiple sentences. The input format for NSP is:

$$\text{[CLS]}\ A\ \text{[SEP]}\ B\ \text{[SEP]},$$

where A and B are sentences, [CLS] is a special token representing the entire sequence, and [SEP] separates sentences.

The NSP objective function is defined as:

$$\mathcal{L}_{\text{NSP}} = -\mathbb{E}_{(A,B)\sim\mathcal{D}} \log P(y|\text{CLS}(A, B)), \tag{4}$$

where y indicates whether B follows A ($y = 1$ for consecutive sentences; $y = 0$ for random pairs).

Examples:

- Positive Pair: A: "Q2 EBITDA margin improved to 22%." B: "This was driven by cost optimization initiatives."
- Negative Pair: A: "Net debt stood at \$4.2B." B: "The CEO emphasized dividend sustainability."

NSP helps BERT capture sentence-level coherence, enabling it to understand relationships between different parts of financial documents.

Both MLM and NSP are combined during pre-training, allowing BERT to develop a deep understanding of language structure and context. This pre-training approach makes BERT highly adaptable for fine-tuning on specific NLP tasks like sentiment analysis.

3.2.3 Financial Text Applications

BERT's bidirectional context modeling proves advantageous for financial NLP tasks:

- **Sentiment Ambiguity Resolution**: Disambiguates terms like 'leverage' (financial vs. operational) through context-aware embeddings.

- **Cross-Sentence Inference**: Links forward-looking statements in earnings calls (e.g., “We expect growth...” in Q&A sections) to management discussion.

- **Transfer Learning**: Enables fine-tuning on small financial datasets, critical given the limited labeled data in finance (Araci, 2019).

3.3 FinBERT: Domain-Specific Financial Sentiment Analysis

Building on the general-purpose language understanding capabilities of BERT, FinBERT (Araci, 2019) was developed to address the unique linguistic patterns and challenges of financial texts. By fine-tuning BERT on a corpus of financial documents, FinBERT enhances sentiment analysis in domains where domain-specific terminology plays a critical role.

3.3.1 Architectural Foundation

FinBERT retains the core architecture of BERT but introduces modifications tailored to financial contexts:

- **Pre-training Corpus**: FinBERT is pre-trained on a corpus of financial texts comprising 4.9 billion tokens, sourced from diverse financial documents such as 10-K filings, earnings call transcripts, and analyst reports. This corpus ensures that FinBERT learns sector-specific language patterns.

- **Tokenization**: The tokenizer is enhanced to handle financial terminology (e.g., ‘EBITDA,’ ‘amortization’) and numerical expressions (e.g., “Q2 FY23,” “15% YoY growth”).

- **Sentiment Classification Layer**: A classification head is added to the pre-trained model to map contextual embeddings to sentiment probabilities (Positive, Neutral, Negative).

The input sequence follows the standard BERT format:

$$\text{Input} = [\text{[CLS]}; w_1, w_2, ..., w_n; \text{[SEP]}],$$

where w_i represents tokenized words, [CLS] denotes the aggregate representation for sentiment classification, and [SEP] separates sentences.

3.3.2 Fine-Tuning Methodology

FinBERT is fine-tuned using supervised learning on annotated financial sentiment datasets.

Datasets Two primary datasets are used for fine-tuning:

1. **Financial PhraseBank** (Malo et al., 2013): Contains 4,840 sentences labeled by domain experts into Positive, Negative, and Neutral categories. Example: "EPS beat consensus by $0.12" → Positive.

2. **FiQA Sentiment Analysis Dataset**: Includes 1,173 news headlines with fine-grained sentiment scores ranging from $[-1, 1]$, enabling nuanced sentiment modeling.

Loss Function The cross-entropy loss function optimizes sentiment classification:

$$\mathcal{L}_{\text{classification}} = -\sum_{c=1}^{C} y_c \log P(y_c),$$

where C is the number of sentiment classes (Positive/Neutral/Negative), y_c is the true label for class c, and $P(y_c)$ is the predicted probability.

Training Protocol The fine-tuning process involves:

- Optimizer: AdamW with a learning rate of 3×10^{-5}.
- Batch Size: 32.
- Early Stopping: Based on validation F1-score.
- Epochs: Typically 3-5 epochs for convergence.

3.3.3 Financial Text Applications

FinBERT addresses several key challenges in financial sentiment analysis:

- **Contextual Negation**: Identifies subtle negations such as "margin expansion despite headwinds" (Positive) versus "margin contraction despite tailwinds" (Negative).
- **Sarcasm/Hedging Detection**: Captures mitigating phrases like "so-called 'growth' strategy" as Negative.
- **Numerical Sensitivity**: Processes quantitative cues (e.g., "guidance raised to 15%" → Positive) through learned embeddings for numerical tokens.

3.3.4 Empirical Validation

FinBERT's performance has been validated on benchmark financial sentiment tasks. Table 1 summarizes its results compared to BERT.

Model	Loss	Accuracy	F1 Score
BERT	0.38	0.85	0.84
FinBERT	**0.37**	**0.86**	0.84

Table 1: Performance With Different Pre-training Strategies
This table presents the performance of FinBERT relative to the original BERT model according to three key metrics; Loss, Accuracy, and F1 Score. It comes from the original FinBERT paper and the results are based on a 10-fold cross validation.

3.4 DeBERTa: Decoding-Enhanced BERT With Disentangled Attention

While FinBERT achieves strong performance in financial sentiment analysis through domain-specific fine-tuning, DeBERTa (He et al., 2021) introduces architectural innovations that address limitations in both BERT and FinBERT. By disentangling content and positional information, DeBERTa improves contextual understanding, making it particularly effective for complex financial documents.

3.4.1 Disentangled Attention Mechanism

Unlike standard transformers that conflate content and position information, DeBERTa explicitly disentangles these components through separate vector representations:

- **Content Matrices**: Capture semantic information independent of position.
- **Position Matrices**: Encode relative positional relationships between tokens.

This separation allows the model to process semantic meaning and positional context as distinct aspects of language, particularly beneficial when analyzing financial texts where both the content of statements and their relative positioning can signal sentiment nuances. In the following, the subscript c in Q_c and K_c denotes the content-based components of the query and key matrices, respectively, as opposed to the positional components. The disentangled attention mechanism is mathematically formalized as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q_c K_c^{'} + Q_c P' + P {K_c}'}{\sqrt{d_k}}\right) V. \tag{5}$$

Where:

- Q_c, K_c represent content-based query and key matrices
- P represents the relative position matrix
- V represents the value matrix
- d_k is the dimensionality of the key vectors

This formulation decomposes attention computation into three components: content-to-content ($Q_c K_c^{'}$), content-to-position ($Q_c P^{'}$), and position-to-content ($P K_c^{'}$) interactions. This granular approach enables more precise modeling of relationships between tokens in financial narratives, where subtleties in phrasing can significantly impact sentiment interpretation.

3.4.2 Enhanced Mask Decoder

DeBERTa enhances BERT's MLM objective by incorporating absolute positional information during the decoding phase. This enhanced mask decoder (EMD) improves token prediction by conditioning on both contextual tokens and their absolute positions.

The training objective can be expressed as:

$$\mathcal{L}_{\text{EMD}} = -\mathbb{E}_{x \sim \mathcal{D}} \sum_{i \in M} \log P(x_i | x_{\setminus M}, \text{Pos}_i). \tag{6}$$

Where:

- M is the set of masked token positions
- $x_{\setminus M}$ represents the input sequence with masked tokens
- Pos_i is the absolute position of the i-th token
- $\mathcal{D}$ is the training corpus

By incorporating absolute positions in the decoder, DeBERTa achieves more accurate token predictions, particularly beneficial for financial statements where the absolute position of information (e.g., early vs. late mention of earnings results) can be semantically significant.

3.4.3 Relative Position Encoding

DeBERTa implements a sophisticated position encoding scheme that captures the relative distances between tokens. Unlike BERT's absolute position embeddings, DeBERTa's relative position encoding directly models token-to-token positional relationships:

$$\text{RelPos}(i, j) = \log(|i - j| + 1) \cdot \text{sign}(i - j). \tag{7}$$

Where i and j are the positions of two tokens in the sequence. This logarithmic scaling:

- Gives higher resolution to nearby tokens
- Reduces the impact of distance for far-apart tokens

- Preserves directional information through the sign function

The relative position encoding is particularly advantageous for processing long financial documents such as earnings calls, where relationships between distant pieces of information (e.g., forward guidance mentioned far from previous performance metrics) need to be captured.

3.4.4 Application to Financial NLP

DeBERTa's architectural improvements make it particularly well-suited for financial sentiment analysis for several reasons:

- **Long-Range Dependencies**: The disentangled attention mechanism better captures relationships between distant parts of earnings calls, such as connections between preliminary statements and subsequent explanations.

- **Contextual Disambiguation**: Financial terms often have context-dependent meanings (e.g., 'volatile' can be positive or negative depending on the context). DeBERTa's enhanced contextual modeling improves disambiguation of such terms.

- **Structural Sensitivity**: Earnings calls follow semi-structured formats where position carries meaning. DeBERTa's explicit handling of positional information helps capture these structural aspects.

Empirical benchmarks such as GLUE and SuperGLUE demonstrate DeBERTa's superior performance over BERT across multiple NLP tasks. In financial NLP applications, these improvements translate to more accurate sentiment classification and topic modeling of earnings transcripts, creating a stronger foundation for investment signal generation.

3.5 NLI

Building upon the capabilities of powerful language models like DeBERTa, we now explore NLI, a fundamental task in natural language understanding that benefits significantly from DeBERTa's enhanced architecture.

NLI studies whether a hypothesis can be inferred from a premise, where both are text sequences. It determines the logical relationship between a pair p (premise) and h (hypothesis). These relationships typically fall into three categories:

- **Entailment**: The hypothesis can be inferred from the premise.
- **Contradiction**: The negation of the hypothesis can be inferred from the premise.
- **Neutral**: Neither entailment nor contradiction holds.

For illustration, consider the following examples:

- **Entailment**:

  Premise: "Two women are hugging each other."

  Hypothesis: "Two women are showing affection."

- **Contradiction**:

  Premise: "The man is running."

  Hypothesis: "The man is sleeping."

- **Neutral**:

  Premise: "The musicians are performing for us."

  Hypothesis: "The musicians are famous."

Mathematically, given a premise p and a hypothesis h, a model f predicts the logical relationship r between them:

$$r = f(p, h), \text{ where } r \in \text{Entailment, Contradiction, Neutral.} \tag{8}$$

Modern NLI systems predominantly employ transformer-based architectures. The standard approach involves encoding the premise and hypothesis pair as a single sequence with special tokens:

$$\text{input} = \text{[CLS]}; p; \text{[SEP]}; h; \text{[SEP]}. \tag{9}$$

This input is then processed through the transformer architecture to produce a classification output. The model function can be formalized as:

$$f(p, h) = \text{softmax}(W \cdot \text{Transformer(input)} + b). \tag{10}$$

Where W and b are learnable parameters, and the output is a probability distribution over the three possible relationships. In this analysis, we use positive, negative, and neutral as the three relationships to get sentiment scores.

As NLP systems continue to evolve, NLI remains a critical component for achieving true language understanding and reasoning capabilities. The superior performance of DeBERTa on NLI tasks is highlighted in He et al. (2021) with tests on empirical benchmarks like MNLI.

4 Portfolio Construction

This section outlines the methodological framework for constructing portfolios based on textual sentiment analysis. First, I detail the computation of sentiment scores using four distinct methodological approaches. Next, I present the aggregation schemes employed to synthesize these scores. Finally, I describe the portfolio formation process and the metrics used to evaluate performance. Table 5 at the end of the section gives an overview of all the portfolios constructed using different combinations of methodologies.

4.1 Sentiment Scoring

First, I utilize DeBERTa within an NLI framework (at two different granularity levels), which produces a continuous sentiment score ranging from -1 to 1. This score corresponds to the probabilistic distribution over positive, negative, and neutral sentiment categories. Next, I apply FinBERT, which generates discrete sentiment values of 1, 0, and -1, representing positive, neutral, and negative sentiment, respectively. Finally, I employ the traditional lexicon-based approach developed by Loughran and McDonald (2011), which provides a continuous sentiment score between -1 and 1, although through a less complex methodology compared to transformer-based models. Table 3, presented at the end of this section, provides a comprehensive summary of the methodologies employed for sentiment scoring and their respective characteristics.

4.1.1 NLI With DeBERTa

I compute NLI sentiment scores at two levels of granularity; at the topic level (denoted NLI-t) and at the transcript level (denoted NLI).

Topic-Level Scores Let $K = 15$ denote the number of predefined topics (see Table 2) compiled in collaboration with a professional equity research analyst. For each atom a_i, two scores are computed:

- **Topic Score** ($ts_{ij} \in [0, 1]$): Represents the likelihood that atom a_i discusses topic t_j, as determined by DeBERTa in an NLI setting.
- **Sentiment Score** ($S_{ij}^{\text{NLI-t}} \in [-1, 1]$): Indicates the polarity associated with topic t_j in atom a_i, inferred using DeBERTa's NLI-based sentiment classification.

Formally, for atom a_i and topic t_j, these scores are defined as:

$$ts_{ij} = f_{\text{DeBERTa}}(a_i, t_j), \quad S_{ij}^{\text{NLI-t}} = g_{\text{DeBERTa}}(a_i, t_j), \tag{11}$$

where f_{DeBERTa} and g_{DeBERTa} denote the topic-scoring and sentiment-scoring functions of the model, respectively.

The topic scores ts_{ij} are obtained using the following NLI prompts:

- **Premise:** "`The news is:`" a_i.
- **Hypothesis:** "`The news contains information about:`" t_j

The sentiment scores $S_{ij}^{\text{NLI-t}}$ are computed using these prompts:

- **Premise:** "`The news is:`" a_i.
- **Positive Hypothesis:** "`The news contains positive information about:`" t_j.

- **Negative Hypothesis:** "`The news contains slightly negative information about:`" t_j.

This approach produces $K = 15$ topic and sentiment scores per atom. To address the model's observed bias towards optimistic sentiment classifications, the negative hypothesis prompt includes the word 'slightly.' This adjustment is empirically validated: for example, when the sentence "the food is bad" is inputted, the model yields a negative score of 0.98 with 'slightly' included, compared to 0.87 without it.

Finally, the overall sentiment score for each topic and atom is computed as the difference between the positive and negative scores:

$$S_{ij}^{\text{NLI-t}} = S_{ij,\text{positive}}^{\text{NLI-t}} - S_{ij,\text{negative}}^{\text{NLI-t}}. \tag{12}$$

Topic	**Frequency (%)**	**Number**
Revenues	38.3	1
Guidance	23.8	2
Organic Growth	8.2	3
Expenditure	6.9	4
Profitability	6.1	5
Margins	4.7	6
Mergers and Acquisitions	4.1	7
Cash Flow	2.6	8
Supply Chains	1.6	9
Macroeconomic Environment	1.1	10
Price Increases	0.9	11
Regulation and Compliance	0.8	12
Shareholder Giveback	0.4	13
Financing/Refinancing	0.3	14
Workforce Expansion/Reduction	0.1	15

Table 2: The Frequency of the Topics Among Atoms
This table presents the fifteen different topics selected and their respective frequency of occurrence in percentage. The frequency is computed as the number of time a topic is selected for an atom (based on its topic score) divided by the total number of atoms.

Transcript-Level Scores The less granular scoring methodology employs a synthesis approach where all individual atoms from the same transcript are merged to form a comprehensive summary representing the entire transcript. This aggregated text is then analyzed using DeBERTa within an

NLI framework to generate a singular sentiment score ($S_{T_i}^{\text{NLI}}$) for the complete transcript T_i, without any topic-level granularity.

4.1.2 FinBERT

Although FinBERT's architecture theoretically supports NLI tasks, adapting it for this framework would require extensive fine-tuning on financial NLI datasets, a process requiring significant computational resources and annotated training data. Given these constraints, I employ FinBERT in its default classification mode, where it assigns one of three labels—Positive, Neutral, or Negative—to each atom. These labels are numerically encoded as 1, 0, and -1 respectively. Unlike the NLI methodology, which produces 15 topic-specific sentiment scores per atom, FinBERT generates a single sentiment score per atom S_i^{FB}, limiting its granularity but preserving computational efficiency.

4.1.3 Lexicon-Based Approach

In the third methodological approach, I adopt a lexicon-based strategy, making use of the Loughran and McDonald (2011) dictionary. First, I compile two separate lists, one for positive and one for negative terms as defined by the dictionary. Next, I convert all the atoms' text to lowercase and split it into individual words. For each atom, I iterate over each word and count the occurrences of words classified as positive or negative. The scaled sentiment score S_i^{MD} for atom a_i is calculated as the difference between the number of positive ($N_{i,\text{positive}}$) and negative ($N_{i,\text{negative}}$) words, normalized by the total word count ($N_{i,\text{words}}$) to account for variations in atom lentgh:

$$S_i^{MD} = \frac{N_{i,\text{positive}} - N_{i,\text{negative}}}{N_{i,\text{words}}}. \tag{13}$$

This normalization facilitates the comparison of sentiment scores across atoms of varying lengths by adjusting for differences in word count. As with the FinBERT approach, the dictionary-based method produces a single sentiment score per atom and does not incorporate topic-specific information.

Characteristic	DeBERTa topic (NLI-t)	DeBERTa transcript (NLI)	FinBERT (FB)	Lexicon (MD)
Model Type	Transformer-based (DeBERTa)	Transformer-based (DeBERTa)	Domain-specific BERT	Dictionary-based
NLI Used	Yes	Yes	No	No
Input Processing	Premise-hypothesis pairs	Premise-hypothesis pairs	Direct text input	Word counting
Topic Awareness	Yes (15 predefined topics)	No	No	No
Scoring Mechanism	$S_{ij}^{\text{NLI-t}} = S_{ij,\text{positive}}^{\text{NLI-t}} - S_{ij,\text{negative}}^{\text{NLI-t}}$	$S_{T_i}^{\text{NLI}}$	$S_i^{\text{FB}} = \begin{cases} 1 & \text{if Positive} \\ 0 & \text{if Neutral} \\ -1 & \text{if Negative} \end{cases}$	$S_i^{\text{MD}} = \frac{N_{i,\text{positive}} - N_{i,\text{negative}}}{N_{i,\text{words}}}$
Score Range	-1 to 1	-1 to 1	-1, 0, 1	-1 to 1
Scores Per Atom	Multiple (one per topic)	Transcript score directly	Single	Single
Domain Adaptation	General model (financial prompts)	General model (financial prompts)	Financial domain	Financial lexicon
Contextual Understanding	High	High	High	None (word-level only)
Complexity	High	Medium	Medium	Low

Table 3: Summary of Sentiment Scoring Methodologies
This table outlines the main characteristics of the four methodologies used to compute the sentiment scores. Each line is a characteristic of the model and each column is a methodology. The different identified features are the model type, the use of NLI, the input processing methodology, the topic awareness, the scoring mechanism, the range of the produced score, the number of scores per atom, the domain adaptation, the contextual understanding, and the complexity.

4.2 Sentiment Score Aggregation

The next step involves aggregating the sentiment scores derived from each methodology into a unified score for each transcript and for each quarter. As summarized in table 3 the NLI topic-level methodology generates multiple sentiment scores $S_{ij}^{\text{NLI-t}}$ per atom due to its topic-specific scoring mechanism. This necessitates a two-stage aggregation process:

1. **Intra-atom aggregation**, to consolidate topic-specific sentiment scores into a single sentiment score per atom;

2. **Inter-atom aggregation**, to aggregate the atoms' sentiment scores into a single sentiment score per transcript.

Contrastingly, scores generated using FinBERT (S_i^{FB}) and the Loughran and McDonald (2011) dictionary (S_i^{MD}) produce a single sentiment score per atom, requiring only the inter-atom aggregation step to achieve transcript-level sentiment quantification. In this section, I detail the methodologies employed at each stage of aggregation.

4.2.1 Intra-Atom Aggregation

I evaluate two distinct intra-atom aggregation methods: the first is referred to as naive aggregation, while the second employs an attention-based aggregation approach.

Naive Aggregation In this method, I select a single topic for each atom. This approach assumes topical exclusivity, where an atom is presumed to primarily discuss the topic for which it exhibits the highest topic score. To associate each atom with a single dominant topic, we apply an argmax selection criterion over topic scores. Formally, for atom a_i, the assigned topic index z_i is determined by:

$$z_i = \underset{j \in \{1,\dots,K\}}{\arg\max}\, ts_{ij}, \tag{14}$$

where ts_{ij} denotes the topic score for atom a_i and topic t_j. The corresponding sentiment score for a_i is $S_{i,z_i}^{\text{NLI-t}}$, reflecting the sentiment score of the dominant topic t_{z_i}.

Attention-Based Aggregation In this approach, no explicit topic selection is performed. Instead, the topic scores are normalized such that, for each atom, they form a probability distribution over all topics. Specifically, the normalized topic score $\tilde{ts}_{ij}$ is computed as

$$\tilde{ts}_{ij} = \frac{ts_{ij}}{\sum_{k=1}^{K} ts_{ik}}, \tag{15}$$

where ts_{ij} denotes the original topic score for atom a_i and topic t_j, K is the total number of topics, and $\tilde{ts}_{ij}$ is the normalized topic score, satisfying $\sum_{j=1}^{K} \tilde{ts}_{ij} = 1$ for each atom a_i.

The sentiment score for each atom is then computed as a weighted average of the topic-specific sentiment scores, using the normalized topic scores as weights:

$$S_i^{\text{NLI-t}} = \sum_{j=1}^{K} S_{ij}^{\text{NLI-t}} \cdot \tilde{ts}_{ij}, \tag{16}$$

where $S_{ij}^{\text{NLI-t}}$ is the sentiment score for atom a_i and topic t_j.

Both aggregation methods described above yield a single sentiment score per atom, thereby requiring one more aggregation step to have a single score per transcript.

4.2.2 Inter-Atom Aggregation

I also explore two methods for the inter-atom aggregation step. The first is referred to as the average aggregation, consisting of a simple average, while the second is an attention-based approach.

Average Aggregation In this method, I compute a single average of the transcript's atoms' sentiment scores:

$$S_{T_i}^{\text{Model}} = \frac{1}{n_i} \sum_{i=1}^{n_i} S_i^{\text{Model}}, \tag{17}$$

where $S_{T_i}^{\text{Model}}$ is the final sentiment score for the transcript T_i, Model is any of the three models (NLI-t, FB, or MD) used to compute the sentiment scores, S_i^{Model} is the sentiment score computed with Model and associated with atom a_i, and n_i is the number of atoms in the transcript.

Attention-Based Aggregation The attention-based aggregation method introduces topic frequency-aware weighting to prioritize prevalent topics while maintaining bounded sentiment scores. Let f_j denote the frequency of occurrence of topic t_j across all atoms in transcript T_i. This approach incorporates quadratic normalization to stabilize sentiment scoring:

$$S_{T_i}^{\text{Model}} = \frac{\sum_{j=1}^{K} f_j \left(\sum_{i=1}^{f_j} S_i^{\text{Model}} \right)}{\sum_{j=1}^{K} f_j^2}, \tag{18}$$

where:

- f_j: Frequency of topic t_j in transcript T_i,
- S_{ij}^{Model}: Sentiment score of i-th atom for topic t_j computed with Model,
- K: Total unique topics.

The squared-normalization approach combines two established principles:

1. **Entropy regularization**: The quadratic denominator aligns with term weighting schemes that mitigate topic dominance through variance stabilization. (Paltoglou and Thelwall, 2010)
2. **Attention normalization**: Extends Slot Attention's value scaling strategies (Krimmel et al., 2024) to frequency-based aggregation.

For example, consider a transcript containing:

- 10 atoms classified under 'Revenue' ($f_1 = 10$),
- 5 atoms under 'Margins' ($f_2 = 5$).

The aggregation becomes:

$$\text{Numerator} = 10 \sum_{i=1}^{10} S_{i1}^{\text{Model}} + 5 \sum_{i=1}^{5} S_{i2}^{\text{Model}},$$

$$\text{Denominator} = 10^2 + 5^2 = 125,$$

$$S_{T_i}^{\text{Model}} = \frac{\text{Numerator}}{125} \in [-1, 1].$$

4.2.3 Topic-Level Aggregation

This final method employs two stages of aggregation as previously described in 4.2, but differs in its application. Initially, I apply the naive aggregation step outlined in 4.2.1 and subsequently filter the atoms to retain only those associated with the five most frequently cited topics (see Table 2). This procedure yields multiple $S_{i,z_i}^{\text{NLI-t}}$ such that :

$$t_{z_i} \in \{\text{`Revenues'}, \text{`Guidance'}, \text{`Organic Growth'}, \text{`Expenditure'}, \text{`Profitability'}\}.$$

Subsequently, for each t_{z_i} in this set, I apply the average aggregation step as described in Section 4.2.2, as follows:

$$S_{T_i,t_{z_i}}^{\text{NLI-t}} = \frac{1}{n_{i,z_i}} \sum_{i=1}^{n_{i,z_i}} S_{i,z_i}^{\text{NLI-t}}, \tag{19}$$

where $S_{i,z_i}^{\text{NLI-t}}$ denotes the sentiment score of the i-th atom for topic t_{z_i}, n_{i,z_i} is the number of atoms associated with topic t_{z_i}, and $S_{T_i,t_{z_i}}^{\text{NLI-t}}$ represents the final sentiment score of transcript T_i with respect to topic t_{z_i}.

For illustration, consider a transcript comprising 15 atoms, with three atoms assigned to each of the five topics identified as the most frequent. In this case, the method produces five sentiment scores for the transcript, one corresponding to each of these topics.

Table 4 summarizes the aggregation methods discussed and gives an overview of their characteristics.

Aggregation Level	**Method**	**Description**	**Mathematical Formulation**	**Applicable Models**
Intra-atom	Naive Aggregation	Selects single topic with highest affinity	$z_i = \underset{j \in \{1,\ldots,K\}}{\arg\max}\, ts_{ij}; S_i^{\text{NLI-t}} = S_{i,z_i}^{\text{NLI-t}}$	NLI-t
Intra-atom	Attention-Based	Weighted average using normalized topic scores	$\tilde{ts}_{ij} = \frac{ts_{ij}}{\sum_{k=1}^{K} ts_{ik}}; S_i^{\text{NLI-t}} = \sum_{j=1}^{K} S_{ij}^{\text{NLI-t}} \cdot \tilde{ts}_{ij}$	NLI-t
Inter-atom	Average	Simple arithmetic mean of atom scores	$S_{T_i}^{\text{Model}} = \frac{1}{n_i} \sum_{i=1}^{n_i} S_i^{\text{Model}}$	NLI-t, FB, MD
Inter-atom	Attention-Based	Topic frequency-weighted average	$S_{T_i}^{\text{Model}} = \frac{\sum_{j=1}^{K} f_j \left(\sum_{i=1}^{f_j} S_i^{\text{Model}} \right)}{\sum_{j=1}^{K} f_j^2}$	NLI-t, FB, MD
Topic-level	Combined	Topic-specific averaging for top 5 topics	$S_{T_i,t_{z_i}}^{\text{NLI-t}} = \frac{1}{n_{i,z_i}} \sum_{i=1}^{n_{i,z_i}} S_{i,z_i}^{\text{NLI-t}}$	NLI-t

Table 4: Overview of Sentiment Score Aggregation Methods and Their Characteristics
This table summarizes the different aggregations approaches employed in this study, specifying the aggregation stage, methodological rationale, formal mathematical expressions, and the corresponding sentiment scoring models to which each method is applicable.

4.3 Long-Only Portfolios

In this section, I detail the methodology employed for constructing sentiment-based long-only portfolios. Additionally, I describe the construction of the long-only benchmark portfolios utilized to evaluate the performance of my proposed strategies.

4.3.1 Portfolio Construction

For the portfolio construction methodology, I employ a systematic approach where, for each quarterly rebalancing period, the ten stocks exhibiting the highest aggregated sentiment scores are selected for inclusion in long-only portfolios. These portfolios implement two distinct weighting methodologies, with their nomenclature following the convention:

L_stage1_stage2_method_weighting

- *L* designates the long position orientation
- *stage1* specifies the intra-atom aggregation method, either *naive* or *attention* as detailed in Section 4.2
- *stage2* indicates the inter-atom aggregation approach: *avg* (average) or *attention* (frequency-weighted)
- *method* denotes the sentiment quantification technique: *NLI* (NLI-t model), *Dict* (MD), *FinBERT* (FB), or specific topical focus (e.g., *'Revenues'*)
- *weighting* represents the capital allocation strategy: *EW* (equal-weight) or *SW* (sentiment-weight)

An exception to this nomenclature convention occurs for the *L_One_Summary_weighting* portfolios, which utilize transcript-level sentiment scores directly without requiring multi-stage aggregation, as discussed in Section 4.1.1.

I employ two distinct weighting schemes to investigate different capital allocation strategies:

- **EW Scheme**: Each constituent asset receives an identical weight calculated as:

$$w_{i,t}^{+,\text{EW}} = \frac{1}{n}, \tag{20}$$

where n denotes the number of selected stocks and the $+$ sign differentiates long weights from short weights.

- **SW Scheme**: This innovative allocation strategy proportionally distributes weights according to relative sentiment scores, formalized as:

$$w_{i,t}^{+,\text{SW}} = \frac{S_{i,t}}{\sum_{i=1}^{n} S_{i,t}}, \tag{21}$$

where $S_{i,t}$ represents the sentiment score for security i at period t. The development and analysis of this weighting methodology constitutes a principal investigative focus of this research, offering potential contributions to asset allocation frameworks.

To avoid any data-snooping bias, I construct portfolios based on data available at the end of the previous time period. The portfolio return at time t, R_t^L, is computed using the returns of the stocks at time t, such that:

$$R_t^{\text{L}} = \sum_{i=1}^{n} w_{i,t-1}^{+} \cdot r_{i,t}, \tag{22}$$

where $r_{i,t}$ is the return of the i-th stock at time t, and $w_{i,t-1}^{+}$ is the weight assigned to the i-th stock based on data up to time $t-1$.

The sentiment analysis procedure was initiated in the first quarter of 2013 (Q1 2013). To ensure compliance with the principle of information availability, portfolio return calculations begin in the second quarter of 2013 (Q2 2013). This approach guarantees that portfolio weights are determined solely on the basis of information observable at the end of the preceding quarter.

4.3.2 Benchmarks

The primary benchmark employs a long-only, equally-weighted strategy across the entire investment universe. This approach follows the empirical framework established by DeMiguel et al. (2009), which demonstrates the comparative effectiveness of naive diversification strategies against capitalization-weighted strategies. The equal-weighted portfolio serves as a robust baseline for evaluating the economic significance of sentiment-driven strategies.

I also implement a momentum-based benchmark portfolio, consisting of long positions in the ten best-performing equities from the preceding quarter. While this quarterly rebalancing frequency diverges from the conventional 12-month momentum factor specification (Moskowitz et al., 2011), it maintains temporal alignment with the sentiment strategy's investment horizon. This design enables direct testing of whether observed outperformance stems from sentiment signals or merely captures short-term momentum effects.

For statistical robustness assessment, I generate an ensemble of 1,000 randomized portfolios through bootstrap resampling. Each synthetic portfolio holds ten randomly selected stocks with equal-weight allocation, following the Monte Carlo methodology proposed by Burns (2004). This bootstrap approach mitigates single-benchmark comparison biases.

4.4 Long-Short Portfolios

In this section, I detail the methodologies employed for constructing sentiment-based long-short portfolios. I explore two different construction methods, namely the traditional one and the expanded one. Subsequently, I describe the benchmark portfolios utilized for comparative analysis.

4.4.1 Portfolio Construction

Traditional Long-Short In extending my analysis to long-short portfolios, I retain the selection criteria for long positions as previously defined. For the short component of the portfolio, I use a similar reasoning; the short portion of the portfolio is made of the 10 stocks with the lowest sentiment scores. Portfolio nomenclature follows the convention:

LS_stage1_stage2_method_weighting

- *LS* designates the long-short orientation
- *stage1* specifies the intra-atom aggregation method per Section 4.2
- *stage2* indicates the inter-atom aggregation approach
- *method* denotes the sentiment quantification technique
- *weighting* represents the capital allocation strategy

The exception for *LS_One_Summary_weighting* portfolios persists, utilizing direct transcript-level scores without multi-stage aggregation as detailed in Section 4.1.1.

Short Position Weighting For SW schemes, short-side weights employ inverse proportional allocation based on raw sentiment scores:

$$w_{i,t}^{-} = \frac{-1}{1 + S_{i,t}}, \tag{23}$$

where $S_{i,t}$ represents the sentiment score for security i at period t. This formulation ensures that negative sentiment scores ($S_{i,t} < 0$) receive larger absolute weights

Final weights undergo normalization to enforce cash neutrality:

$$\sum_{i=1}^{n^{-}} w_{i,t}^{-} = -1, \tag{24}$$

where n^{-} is the number of stocks with negative sentiment. The EW scheme maintains symmetric allocation:

$$w_{i,t}^{\pm,\text{EW}} = \pm\frac{1}{n}. \tag{25}$$

The EW methodology ensures portfolios are cash-neutral, with both portions of the portfolio (long and short) summing to either 1 or -1.

Expanded Long-Short I develop an alternative portfolio construction approach, the expanded long-short strategy, which introduces conditional short allocation based on explicit negative sentiment signals, denoted as *LS_stage1_stage2_method_exp_weighting*. This adaptive framework employs:

- **Short Position Criteria**: Exclusively targets equities with negative sentiment scores ($S_{i,t} < 0$)
- **Positioning Logic**:

$$w_{i,t}^{-} = \begin{cases} \dfrac{-1}{n^{-}} & \text{For EW} \\ \dfrac{-S_{i,t}}{\sum_{i=1}^{n^{-}} S_{i,t}} & \text{For SW} \end{cases} \tag{26}$$

  where n^{-} denotes the number of negative sentiment stocks (capped at 10), and $S_{i,t}$ represents negative sentiment scores.

- **Fallback Mechanism**: Implements long-only positioning ($\sum w_{i,t} = 1$) when $n^{-} = 0$, selecting the ten highest-scoring equities with equivalent weighting scheme

This methodology enables:

1. Direct testing of negative sentiment signal efficacy through explicit short positioning
2. Empirical distinction between weakly positive and explicitly negative classifications

The long portion maintains identical weighting logic to standard long-short strategies:

$$w_{i,t}^{+} = \begin{cases} \dfrac{1}{n^{+}} & \text{For EW} \\ \dfrac{S_{i,t}}{\sum_{j=1}^{n^{+}} S_{j,t}} & \text{For SW} \end{cases} \tag{27}$$

where n^{+} denotes the number of long positions (fixed at 10).

While deviating from long short when $n^{-} = 0$, this approach enables direct evaluation of whether negative sentiment scores contain predictive information distinct from merely attenuated positive signals. The performance differential between standard and expanded weighting strategies offers insights into the model's capacity to identify both outperformance and underperformance candidates.

The combined long-short portfolio return at time t, R_t^{LS}, is obtained by summing the products of the returns and weights of the long and short positions:

$$R_t^{\mathrm{LS}} = \sum_{i=1}^{n^+} w_{i,t-1}^+ \cdot r_{i,t} + \sum_{j=1}^{n^-} w_{j,t-1}^- \cdot r_{j,t}, \tag{28}$$

where $w_{i,t-1}^+$ and $w_{j,t-1}^-$ are the weights assigned to the long and short positions, respectively, based on data available up to time $t-1$, and $r_{i,t}$ and $r_{j,t}$ are the returns for the corresponding long and short stocks at time t. Through this method, I assess the performance of my methodology in both long and short investment decisions.

4.4.2 Benchmarks

For the long-short framework, I do not construct a simple benchmark analogous to the long-only equal-weighted portfolio over the entire investment universe. As detailed in Section 2.1, the temporal variability in stock availability—where constituents enter or exit the universe over time—would necessitate arbitrary selection mechanisms to choose long and short positions. Instead, I generate 1,000 random long-short portfolios, consistent with the methodology employed in the long-only setting. In each portfolio, both the long and short legs consist of 10 randomly selected stocks from the investment universe.

Additionally, I apply the same rationale for the momentum benchmark as in the long-only framework. Specifically, I construct a portfolio that takes long positions in the 10 stocks with the highest returns over the previous quarter and short positions in the 10 stocks with the lowest returns over the same period.

	NLI-t	MD	NLI (One Summary)	FB	Topic	Total
Long-Only	6	4	2	4	10	26
Traditional Long-Short	6	4	2	4	10	26
Expanded Long-Short	6	4	2	4	10	26
Total	18	12	6	12	30	78

Table 5: Total Number of Portfolios per Methods
This table provides an overview of the number of portfolios generated by each sentiment scoring methodology and portfolio construction framework. For the NLI-t portfolios, two stages of aggregation are employed, resulting in three possible combinations for three distinct sentiment scores. Each of these scores is subsequently used to construct portfolios with both EW and SW schemes, yielding a total of six portfolios. The MD and FB methods involve only a single aggregation stage, producing two possible combinations for each sentiment score, and thus four portfolios in total (two for EW and two for SW). The simple NLI approach, which applies sentiment scoring to a single summary without aggregation, results in one portfolio for EW and one for SW. Finally, as five topics are selected for the topic-based scoring approach, this leads to the construction of ten portfolios (five for EW and five for SW).

4.5 Performance Measures

In my portfolio performance evaluation, I utilize a diverse set of metrics that encompass both absolute returns and risk-adjusted measures, as well as other risk indicators. My comparison is based on seven key metrics: cumulative gains, annualized mean return, annualized standard deviation, annualized Sharpe ratio, maximum drawdown (MDD), annualized alpha, and annualized turnover.

The cumulative returns, annualized mean returns, and annualized standard deviation are straight forward standard measures, so I do not go over their methodology in details.

Sharpe Ratio The Sharpe ratio is a metric for assessing risk-adjusted returns by quantifying excess return per unit of volatility. To address temporal variations in the risk-free rate, this analysis implements a rolling-window methodology. I use quarterly portfolio returns to compute the Sharpe ratios beginning in Q2 2014. The 4 quarter rolling period allows me to incorporate a full year of data for each Sharpe ratio calculation. The Sharpe ratio for each quarter is defined as:

$$SR_{Q_i} = \frac{r_{Q_i} - r_{f_{Q_i}}}{\bar{\sigma}_j}. \tag{29}$$

The Sharpe ratio is then multiplied by 2 for annualization. Finally, the overall average annualized Sharpe ratio is computed by averaging these annualized values obtained from each rolling window:

$$\bar{SR}_A = \frac{1}{N}\sum_{i=1}^{N} SR_{A_i}, \tag{30}$$

where:

- SR_{Q_i} is the Sharpe ratio for the i-th quarter, starting at the fourth one,
- r_{Q_i} is the return of the portfolio for the i-th quarter,
- $r_{f_{Q_i}}$ is the risk-free rate corresponding to the i-th quarter,
- $\bar{\sigma}_j$ is the standard deviation of the portfolio's returns over the j-th rolling window,
- SR_{A_i} is the annualized Sharpe ratio for the i-th quarter, and
- $\bar{SR}_A$ is the average of the annualized Sharpe ratios over the whole time period, with N representing the total number of such windows available for analysis.

This methodology stems from the findings of Tang and Whitelaw (2011) who demonstrate that Sharpe ratios are low at the peak of a cycle and high at the trough. Thus, computing a rolling sharpe ratio provides a more representative measure of risk-adjusted returns taking into account short term fluctuations.

MDD The MDD is a measure of the peak-to-trough decline of an investment portfolio. It is mathematically defined as the maximum observed loss from a peak to a trough of a portfolio, before a new peak is attained. The MDD is expressed as a percentage and is computed as:

$$\text{MDD} = \max_{\tau\in[0,T]}\left(\max_{t\in[0,\tau]}(CR_t) - CR_\tau\right), \tag{31}$$

$$\text{MDD (\%)} = \left(\frac{\text{MDD}}{\max_{t\in[0,T]}(CR_t)}\right) \times 100, \tag{32}$$

where:

- CR_t is the cumulative return at time t,

- T is the total period under consideration,

- $\max_{t\in[0,\tau]}(CR_t)$ is the running maximum cumulative return up to time τ,

- $\max_{\tau\in[0,T]}$ is the operation of finding the maximum drawdown over the total period T.

Alpha To estimate alpha, I conduct a regression of the portfolio returns on the Fama-French six-factor model which comprises the market excess returns ($R_m - R_f$), size (SMB), value (HML), profitability (RMW), investment (CMA), and momentum (MOM) factors. The model is specified as follows:

$$\begin{aligned} R_p - R_f = \alpha + \beta_m(R_m - R_f) + \beta_{SMB}SMB + \beta_{HML}HML \\ + \beta_{RMW}RMW + \beta_{CMA}CMA + \beta_{MOM}MOM + \epsilon, \end{aligned} \tag{33}$$

where R_p denotes the portfolio return, R_f is the risk-free rate, and ϵ is the error term. Then, I observe the α coefficient and its associated p-value to assess the statistical significance of this parameter. To ensure the robustness of the inference and to address potential issues arising from heteroskedasticity, I estimate the OLS regression using Newey-West standard errors.

Turnover To quantify portfolio rebalancing activity, I calculate quarterly turnover as the sum of absolute weight changes across all holdings. Mathematically, $w_{i,t}^{\text{after-rebalancing}}$ represents the weight of asset i in the portfolio at time t right after rebalancing the portfolio, and $w_{i,t}^{\text{before-rebalancing}}$ represents the weight of the same asset right before the portfolio is rebalanced. The turnover for asset i is given by $|w_{i,t}^{\text{after-rebalancing}} - w_{i,t}^{\text{before-rebalancing}}|$. Therefore, the total turnover of the portfolio is computed as:

$$\text{Turnover}_t = \frac{1}{2} \times \sum_i |w_{i,t}^{\text{after-rebalancing}} - w_{i,t}^{\text{before-rebalancing}}| \times 4. \tag{34}$$

Here, the summation is taken over all assets i in the portfolio. The factor of $\frac{1}{2}$ is included to account for the fact that turnover measures total buying and selling activity; without this adjustment, turnover would be double-counted. The final turnover figure is annualized by multiplying by 4.

5 Empirical Results

In this section, I divide my experiments into two main categories, each further segmented into three sub-categories for clarity. The two primary categories are Long-Only Portfolios and Long-Short Portfolios. The sub-categories are as follows:

The first sub-category, Granular Sentiment-Aggregated Portfolios, includes portfolios constructed using the *naive_avg*, *naive_attention*, and *attention_avg* methodologies. The second sub-category is Topic Portfolios, which contains the portfolios built on individual topics, and the final sub-category is One-Summary Portfolios, containing the results on the portfolio constructed with one comprehensive summary per transcript.

By structuring the analysis in this way, I aim to provide a comprehensive understanding of the different portfolios evaluated in this study. Table 5 shows all the different portfolios constructed in this thesis.

5.1 Long-Only Portfolios

5.1.1 Granular Sentiment-Aggregated Portfolios

Based on the results presented in Table 6, several key insights emerge regarding the influence of weighting methods, sentiment calculation techniques, and sentiment aggregation strategies on portfolio performance. Each of these dimensions contribute distinctively to the observed portfolio metrics. Figure 1 presents the cumulative returns of the Naive-Average portfolios, demonstrating the outperformance of those strategies relative to the benchmark and highlighting that my portfolios are positioned in the upper percentiles compared to the 1,000 generated random portfolios. Graphs for other sentiment aggregation methods are provided in the Appendix.

Overall, the SW method consistently outperforms the EW method across most metrics. For example for the naive-average DeBERTa portfolios, the SW method yields cumulative gains of 18.35$ and mean return of 34.9% whereas the EW method yields 17.16$ in cumulative gains, and 34.1% in mean return. The only metrics where the EW method shows an advantage are volatility and turnover. This outcome is intuitive, as the SW method may require rebalancing even without changes in the portfolio composition, driven by fluctuations in sentiment scores and prices. In contrast, the EW portfolio is driven by fluctuations in prices only. Using the same example as before, the EW scheme yields volatility of 26.1% and turnover of 242% whereas the SW one yields volatility of 26.3% and turnover of 247%.

Regarding sentiment calculation methods, the results are clear: portfolios based on the NLI-t method outperform those based on the MD and FB approaches. The NLI-t portfolios show higher cumulative gains (17.16$ for EW and 18.35$ for SW) and a better Sharpe ratio (2.36 for EW and 2.21 for SW) while maintaining lower volatility, maximum drawdown, and turnover. These findings suggest that the sentiment signals generated by DeBERTa are more effective at identifying favorable market conditions, likely due to the model's advanced language processing capabilities. FinBERT also performs well, exhibiting slightly higher mean returns (35.6% for EW and 35.8% for SW) and alpha (6.4% for EW and 6.5% for SW) compared to DeBERTa (5.8% for EW and 6.0% for SW). This indicates that FB portfolios may deliver more consistent returns, while NLI-t portfolios demonstrate greater variability, with a return distribution characterized by fatter tails. The dictionary-based approach, while simpler, generally underperforms both the NLI-t and FB methodologies.

Lastly, the naive-average aggregation method outperforms other techniques across most metrics, except for turnover and maximum drawdown. By analyzing the differences in holdings between the naive-average and naive-attention portfolios—the two best-performing aggregation methods in the long-only setting—I find an average deviation of 4, or 40% of the portfolio. This highlights a significant difference in the final sentiment scores produced by these two methods. The Naive-Attention method results in slightly lower cumulative returns and higher volatility, suggesting that while it is more responsive to frequently discussed topics, it may also amplify noise or less relevant sentiment shifts. In contrast, the attention-average method underperforms both, indicating that while it seeks to moderate the influence of certain topics, this balancing act may dilute the effectiveness of sentiment signals.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Naive-Average Equally-Weighted							
DeBERTa (NLI-t)	17.16	0.341	**0.261**	**2.36**	43.11	0.058^{*}	2.42
FinBERT (FB)	17.04	0.356	0.319	1.34	52.82	0.064^{*}	2.81
Dict (MD)	10.79	0.294	0.296	1.44	50.30	0.048^{**}	2.50
Panel B: Naive-Average Sentiment-Weighted							
DeBERTa	**18.35**	0.349	0.263	2.21	41.71	0.06^{*}	2.47
FinBERT	17.29	**0.358**	0.320	1.32	52.94	$\mathbf{0.065}^{*}$	2.83
Dict	11.76	0.306	0.304	1.48	49.65	0.05^{**}	2.58
Panel C: Naive-Attention Equally-Weighted							
DeBERTa	16.68	0.344	0.286	1.84	41.36	0.057^{*}	2.42
FinBERT	15.49	0.343	0.311	1.11	49.07	0.059^{*}	2.78
Dict	13.19	0.330	0.329	1.64	56.77	0.059^{**}	2.63
Panel D: Naive-Attention Sentiment-Weighted							
DeBERTa	18.25	0.355	0.289	1.80	**40.78**	0.059^{*}	2.48
FinBERT	15.59	0.344	0.312	1.11	49.23	0.06^{*}	2.81
Dict	13.71	0.334	0.332	1.68	55.69	0.06^{**}	2.71
Panel E: Attention-Average with Both Weighting Methods							
DeBERTa EW	16.44	0.34	0.28	1.83	46.90	0.058^{*}	**2.38**
DeBERTa SW	17.59	0.35	0.29	1.83	45.65	0.059^{*}	2.44

Table 6: Descriptive Statistics for Long-Only Granular Portfolios (Annualized)
This table presents annualized performance metrics for long-only portfolios constructed using various sentiment extraction methodologies and portfolio construction frameworks. The sentiment methodologies include DeBERTa (NLI-t), FinBERT (FB), and the dictionary-based approach (MD). Portfolios are evaluated under different weighting schemes—EW and SW—and aggregation strategies, including Naive-Average, Naive-Attention, and Attention-Average. The reported metrics are cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and portfolio turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** at the 95% level.
Bold values highlight the best-performing metric within each panel.

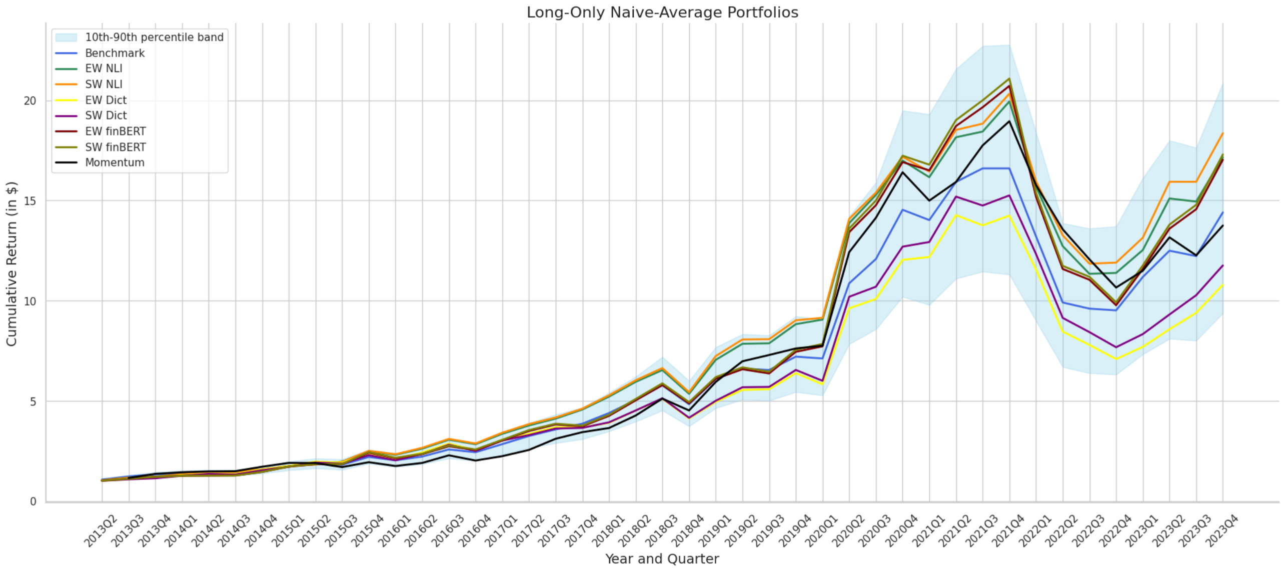

Figure 1: Cumulative Returns of Long-Only Naive-Average Portfolios
This figure displays the cumulative returns of long-only portfolios constructed using the naive-average sentiment aggregation method, evaluated under both EW and SW schemes. The results are shown for each sentiment extraction methodology (NLI, FinBERT, and Dictionary), alongside the benchmark, a momentum strategy, and the 10th–90th percentile band of randomly generated portfolios. The x-axis represents the evaluation period by quarter, and the y-axis reports the cumulative return in dollars.

5.1.2 Topic Portfolios

The results presented in Table 7 demonstrate that portfolio performance varies significantly depending on the topics analyzed, highlighting the importance of specific sentiment themes in driving returns. This more detailed analysis is conducted to pinpoint which topics contribute most to the outperformance observed in the previous aggregation methods.

In terms of weighting methodologies, I generally observe the same trends as before, with SW outperforming EW. Regarding the topics themselves, I find that 'Expenditure' and 'Profitability' yield the highest returns. This result may seem counterintuitive for 'Expenditure,' as it is not a direct performance metric. However, these results must be interpreted in light of the frequency with which each topic was mentioned. For instance, the 'Revenue' topic was mentioned 38.3% of the time, as shown in Table 2. A higher frequency means that there is a higher chance of having false positives (sentiment related to this topic wrongly labeled). In other words, the frequency of topic occurrences must be factored in when evaluating the results; the higher the frequency, the greater the likelihood of model errors, which can affect portfolio performance. Since 'Profitability' and 'Expenditure' are the two least frequently mentioned topics among the five selected for portfolio construction, further investigation is needed to determine whether these topics are indeed the most influential in driving outperformance.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Equally-Weighted							
Revenue	12.04	0.301	0.275	1.22	50.30	0.054*	**2.64**
Guidance	15.33	0.338	0.299	1.20	47.4	0.056*	2.77
Growth	8.11	0.253	0.265	1.28	53.3	0.041**	2.67
Expenditure	18.60	0.366	0.325	1.95	44.39	0.065*	2.92
Profitability	**22.20**	**0.381**	0.297	2.07	47.26	**0.074***	2.90
Panel B: Sentiment-Weighted							
Revenue	12.20	0.303	0.278	1.23	50.33	0.055*	2.69
Guidance	16.53	0.343	0.286	1.18	44.35	0.060*	2.93
Growth	8.75	0.261	**0.262**	1.31	49.63	0.042**	2.86
Expenditure	19.41	0.372	0.334	1.51	**38.75**	0.065*	3.41
Profitability	20.51	0.374	0.305	**2.72**	49.76	0.071*	3.21

Table 7: Descriptive Statistics for Long-Only Topic Portfolios (Annualized)
This table reports annualized performance metrics for long-only portfolios constructed based on sentiment scores associated with five key topics (Revenue, Guidance, Growth, Expenditure, and Profitability) extracted from earnings call transcripts. Results are presented for both equally-weighted (Panel A) and sentiment-weighted (Panel B) portfolio construction methods. The reported statistics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

5.1.3 One-Summary Portfolios

To assess whether the high level of granularity in my sentiment extraction truly enhances portfolio performance, I analyze the outcomes of the One-Summary portfolios. These portfolios are constructed using sentiment extracted from a single summary of each transcript, as opposed to the more granular, atom-level sentiment extraction applied in previous analyses.

The results, presented in Table 8, clearly show that the performance of the One-Summary portfolios is significantly lower across all metrics compared to the granular sentiment-based portfolios discussed before. Both the EW and SW One-Summary portfolios exhibit diminished performance, indicating that relying on a single summary sentiment lacks the nuanced insights that drive higher returns and better risk-adjusted performance. Additionally, maximum drawdowns are higher compared to the granular sentiment-based portfolios, suggesting that the One-Summary approach is less effective in capturing sentiment shifts that could mitigate downside risks. These findings reinforce the conclusion that a more granular approach to sentiment extraction provides substantial value to portfolio performance.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Equally-Weighted	9.74	0.288	0.312	**1.22**	52.69	0.052^{**}	**2.73**
Sentiment-Weighted	**10.40**	**0.295**	**0.311**	1.20	**52.57**	$\mathbf{0.054}^{**}$	2.84

Table 8: Descriptive Statistics for Long-Only One-Summary Portfolios (Annualized)
This table reports annualized performance metrics for long-only portfolios constructed using sentiment extracted from a single summary of each earnings call transcript. Results are presented for both equally-weighted and sentiment-weighted portfolio construction methods. The reported statistics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

5.2 Long-Short Portfolios

The results of the long-short portfolios reveal some interesting insights. Unlike the long-only portfolios, the long-short configurations exhibit significantly weaker performance across most metrics. This suggests that while sentiment-driven strategies may be effective in long-only contexts, their impact in long-short portfolios could be more limited. However, Schuettler et al. (2024) demonstrate that their model performs better in a long-short setting, indicating that my findings should not be generalized to all sentiment analysis models. The weaker performance in my case may be attributed to the limited universe of 69 stocks within the software sector, which has seen exceptional performance over the past 10 years, as shown in Table 14 in the appendix. Such strong performance makes it challenging to generate positive returns using a short strategy. The only notable advantage of the long-short strategies compared to the long-only portfolios is for risk management purposes. Although the Sharpe ratios are considerably lower due to weaker returns, I observe significantly lower volatility and maximum drawdowns.

In this section, I analyze the results from both the expanded and regular portfolios to provide a more comprehensive perspective. However, I believe the expanded strategies offer a more accurate reflection of my model's performance. These results clearly demonstrate the model's ability to assign negative sentiment scores to the appropriate stocks, whereas the regular framework may result in shorting stocks with positive but lower sentiment scores.

5.2.1 Granular Sentiment-Aggregated Portfolios

The results from the regular strategy are shown in Table 9, while those from the expanded strategies are shown in Table 10.

Regular Strategies The trends observed earlier are less pronounced in this section. While the SW methodology remains the strongest compared to EW, and dictionary-based portfolios continue to underperform the other two, DeBERTa generally underperforms relative to FinBERT. Specifically, DeBERTa outperforms FinBERT in the naive-attention setting but lags behind in the naive-average configuration. The best-performing portfolio across all metrics is the FinBERT-based portfolio in the SW naive-average context.

Expanded Strategies I draw the same conclusions for the expanded strategies regarding the comparison between DeBERTa and FinBERT, with the exception of the naive-attention SW setting. FinBERT portfolios also demonstrate more consistent performance than DeBERTa, indicating that regardless of the aggregation method, FinBERT tends to perform well across the board. However, in this case, the DeBERTa SW naive-attention portfolio delivers the best performance and is the only portfolio among all my long-short strategies to generate significant alpha.

While the performance gap between long-short and long-only strategies might be partially explained by my investment universe, these results clearly indicate that DeBERTa, used in an NLI framework, struggles to capture negative sentiment and underperforms FinBERT. This outcome is not surprising, as companies rarely make explicit negative statements. Prior research has shown that complexity in financial language is often correlated with poor performance. Therefore, FinBERT, being specifically trained on financial data, may be better equipped to capture the subtleties of financial language that signal negative sentiment.

It is difficult to draw definitive conclusions from these mixed results, but they do highlight some of the limitations of my model and the investment universe used in this analysis.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Naive-Average Equally-Weighted							
DeBERTa (NLI-t)	1.55	0.051	**0.138**	0.61	26.22	0.006	2.37
FinBERT (FB)	1.56	0.056	0.166	0.65	35.19	0.007	2.71
Dict (MD)	0.99	0.014	0.175	-0.07	38.37	-0.005	2.44
Panel B: Naive-Average Sentiment-Weighted							
DeBERTa	1.58	0.068	0.226	-0.06	21.66	0.013	2.67
FinBERT	**2.69**	0.116	0.200	**0.68**	**16.84**	**0.018**	2.96
Dict	1.42	0.083	0.323	0.12	42.27	-0.004	2.81
Panel C: Naive-Attention Equally-Weighted							
DeBERTa	1.84	0.069	0.145	0.11	26.66	0.011	2.39
FinBERT	1.54	0.052	0.146	0.28	33.67	0.003	2.70
Dict	1.19	0.032	0.182	-0.06	29.36	-0.001	2.51
Panel D: Naive-Attention Sentiment-Weighted							
DeBERTa	1.01	0.036	0.271	-0.26	44.76	-0.008	2.69
FinBERT	1.92	0.079	0.182	0.23	33.81	0.011	2.96
Dict	2.15	**0.119**	0.314	0.11	42.57	0.011	2.83
Panel E: Attention-Average with Both Weighting Methods							
DeBERTa EW	1.19	0.028	0.154	0.17	29.98	0.002	2.35
DeBERTa SW	1.23	0.042	0.220	0.04	47.13	0.007	2.67

Table 9: Descriptive Statistics for Regular Long-Short Granular Portfolios (Annualized)
This table reports annualized performance metrics for regular long-short portfolios constructed using granular sentiment scores derived from earnings call transcripts. Portfolios are formed using three sentiment extraction methodologies: DeBERTa, FinBERT, and the dictionary-based approach, and are evaluated under different aggregation schemes (Naive-Average, Naive-Attention, and Attention-Average) and weighting methods (equally-weighted and sentiment-weighted). The reported metrics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Naive-Average Equally-Weighted							
DeBERTa (NLI-t)	2.94	0.150	**0.292**	0.82	48.63	0.030	2.94
FinBERT (FB)	**6.77**	**0.246**	0.310	0.93	52.82	0.043	3.00
Dict (MD)	2.41	0.147	0.347	0.60	61.70	0.013	3.15
Panel B: Naive-Average Sentiment-Weighted							
DeBERTa	3.30	0.163	0.294	0.97	46.89	0.034	2.95
FinBERT	6.63	0.244	0.312	0.88	52.94	0.042	3.00
Dict	2.56	0.156	0.357	0.64	61.28	0.014	3.15
Panel C: Naive-Attention Equally-Weighted							
DeBERTa	3.75	0.180	0.305	0.59	36.33	0.028	2.99
FinBERT	4.38	0.199	0.315	0.64	53.50	0.036	3.10
Dict	2.05	0.127	0.341	0.67	60.60	0.010	3.21
Panel D: Naive-Attention Sentiment-Weighted							
DeBERTa	6.87	0.250	0.314	**1.08**	**30.34**	**0.051****	3.07
FinBERT	4.71	0.208	0.316	0.67	53.65	0.038	3.11
Dict	1.63	0.108	0.356	0.31	64.10	0.001	3.23
Panel E: Attention-Average with Both Weighting Methods							
DeBERTa EW	1.84	0.106	0.308	0.33	55.20	0.019	**2.89**
DeBERTa SW	2.15	0.122	0.311	0.29	53.92	0.024	**2.89**

Table 10: Descriptive Statistics for Expanded Long-Short Granular Portfolios (Annualized)
This table reports annualized performance metrics for expanded long-short portfolios constructed using granular sentiment scores derived from earnings call transcripts. Portfolios are formed using three sentiment extraction methodologies: DeBERTa, FinBERT, and the dictionary-based approach, and are evaluated under different aggregation schemes (Naive-Average, Naive-Attention, and Attention-Average) and weighting methods (equally-weighted and sentiment-weighted). The expanded long-short framework selects short positions exclusively among stocks with negative sentiment scores, allowing for a variable number of short positions each quarter. The reported metrics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

5.2.2 Topic Portfolios

The performance of the long-short topic portfolios presents a similarly challenging picture. The results in Table 11 correspond to the regular long-short topic portfolios, while those in Table 12 are from the expanded portfolios.

Regular Strategies The regular strategies show consistent results, with SW portfolios outperforming the EW ones. However, when comparing the top-performing topics with the long-only setting, there is a notable shift from ‘Expenditure’ and ‘Profitability’ to ‘Guidance’ and ‘Expenditure.’ ‘Guidance’ can be tricky to interpret, as it is often subject to varying expectations (e.g., whether guidance meets or exceeds analyst forecasts). Nonetheless, it emerges as the best-performing portfolio in both absolute and risk-adjusted metrics. ‘Expenditure’ retains its significance, as it also features prominently in the long-only strategies.

Expanded Strategies In the expanded strategies, an interesting development occurs. Firstly, SW does not consistently outperform EW. Additionally, ‘Expenditure’ shows lower performance and is no longer among the top two topics, being replaced by ‘Growth.’ This is an intriguing finding, indicating that the strong performance of ‘Guidance’ in the regular strategies is likely tied to the long positions in the portfolio. In the expanded strategy, where only negative sentiment stocks are shorted, shorting stocks associated with negative ‘Expenditure’ sentiment performs poorly. The strong performance of the ‘Growth’ portfolios suggests that this topic is more frequently linked to negative sentiment, allowing greater profit capture through short positions. Furthermore, the performance of the ‘Guidance’ portfolios improves compared to the regular strategy, demonstrating that the model effectively assigns negative sentiment to guidance-related discussions.

These findings reinforce my intuition that the expanded methodology provides a more accurate assessment of the model’s ability to identify and label negative sentiment.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Equally-Weighted							
Revenue	0.61	-0.031	0.168	-0.26	49.90	-0.007	**2.70**
Guidance	1.78	0.067	0.147	0.42	**22.65**	0.006	2.85
Growth	0.69	-0.025	**0.137**	-0.18	45.10	-0.017	2.81
Expenditure	0.98	0.012	0.174	-0.36	42.16	-0.012	2.99
Profitability	0.76	-0.005	0.208	-0.40	52.13	0.002	2.98
Panel B: Sentiment-Weighted							
Revenue	0.90	0.025	0.256	0.08	68.02	0.009	2.91
Guidance	**2.88**	**0.127**	0.213	**0.91**	33.69	**0.018**	2.99
Growth	0.72	-0.016	0.173	-0.14	47.18	-0.022	3.15
Expenditure	1.78	0.100	0.315	-0.27	60.21	0.014	3.37
Profitability	0.39	-0.020	0.339	0.46	76.42	0.001	3.37

Table 11: Descriptive Statistics for Regular Long-Short Topic Portfolios (Annualized)
This table presents annualized performance metrics for regular long-short portfolios constructed based on sentiment scores associated with five key topics (Revenue, Guidance, Growth, Expenditure, and Profitability). Results are stratified into equally-weighted (Panel A) and sentiment-weighted (Panel B) construction methods. Reported metrics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Equally-Weighted							
Revenue	0.76	0.061	0.369	0.54	86.52	0.029	3.03
Guidance	2.53	0.107	**0.180**	0.42	**28.07**	0.015	3.16
Growth	5.18	0.205	0.275	**0.99**	49.03	0.022	**3.00**
Expenditure	1.35	0.058	0.26	-0.14	46.06	0.010	3.20
Profitability	1.00	0.085	0.379	0.69	76.24	0.024	3.37
Panel B: Sentiment-Weighted							
Revenue	0.57	0.071	0.400	0.55	91.18	**0.037**	3.08
Guidance	3.24	0.140	0.213	0.67	34.21	0.019	3.30
Growth	**5.61**	**0.210**	0.263	0.76	41.53	0.024	3.11
Expenditure	0.93	0.033	0.302	-0.13	56.10	-0.005	3.38
Profitability	0.96	0.082	0.381	0.70	78.48	0.021	3.48

Table 12: Descriptive Statistics for Expanded Long-Short Topic Portfolios (Annualized)
This table presents annualized performance metrics for expanded long-short portfolios constructed using sentiment scores associated with five key topics (Revenue, Guidance, Growth, Expenditure, and Profitability). Results are stratified into equally-weighted (Panel A) and sentiment-weighted (Panel B) construction methods within the expanded framework that exclusively shorts stocks with negative sentiment scores. Reported metrics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

5.2.3 One-Summary Portfolios

The results for the One-Summary portfolios are presented in Table 13. There is little to elaborate on here, as these portfolios consistently show poor performance across both long-only and long-short settings. The cumulative returns indicate that, in the regular strategies, the portfolio fails to add value, ultimately finishing with a cumulative return of 1. In the expanded strategies, it even destroys value, with a cumulative return falling below 1. These results further confirm the relevance of my granular analysis, demonstrating that relying solely on a comprehensive summary of the transcript is insufficient to capture both positive and negative sentiment effectively.

Portfolio	Cum Gain	Mean Return	Std	Sharpe	Max DD (%)	Alpha	Turnover
Panel A: Regular strategies							
Equally-Weighted	**1.00**	0.010	**0.141**	**-0.13**	**34.18**	0.003	**2.66**
Sentiment-Weighted	**1.00**	**0.025**	0.231	-0.32	37.09	**0.006**	2.97
Panel B: Expanded strategies							
Equally-Weighted	0.57	0.001	0.335	-0.24	70.37	0.000	3.10
Sentiment-Weighted	0.45	-0.014	0.349	-0.23	76.79	-0.003	3.16

Table 13: Descriptive Statistics for Long-Short One-Summary Portfolios (Annualized)
This table presents annualized performance metrics for long-short portfolios constructed using sentiment scores derived from single summaries of earnings call transcripts. Results are stratified into regular strategies (Panel A) and expanded strategies (Panel B), with both equally-weighted and sentiment-weighted construction methods. Reported metrics include cumulative gain, mean return, standard deviation (volatility), Sharpe ratio, maximum drawdown (in percentage), alpha, and annualized turnover. Significance levels are denoted as follows:
* indicates significance at the 99% level,
** indicates significance at the 95% level.
Bold values highlight the best-performing metric within each panel.

5.3 Summary of Empirical Results

Long-Only Portfolios

- **Weighting Scheme Superiority:** The sentiment-weighted methodology consistently outperforms equally-weighted strategies across key performance metrics, including cumulative gains and risk-adjusted returns.
- **Model Efficacy:** The DeBERTa model demonstrates superior capability in capturing positive sentiment signals within earnings call transcripts when deployed in an NLI framework at the

atom level.

- **Granularity Impact:** Granular sentiment extraction significantly enhances portfolio performance, as evidenced by the underperformance of One-Summary portfolios relative to methodologies employing atom-level sentiment scoring.

- **Topic-Specific Contributions:** Profitability and Expenditure topics exhibit the strongest association with positive sentiment-driven returns. However, these findings may reflect sampling bias due to the lower frequency of these topics in earnings calls (see Table 2).

- **Aggregation Method Dominance:** The Naive-Average aggregation approach delivers superior absolute and risk-adjusted returns compared to attention-based alternatives.

Long-Short Portfolios

- **Negative Sentiment Detection:** Preliminary evidence suggests FinBERT may outperform DeBERTa in identifying actionable negative sentiment signals, though results remain statistically inconclusive.

- **Context-Dependent Aggregation:** While Naive-Average remains effective in long-only frameworks, its superiority diminishes in long-short strategies, with no single aggregation method demonstrating consistent dominance.

Consistent Drivers

- **Granularity:** Granular sentiment extraction retains its critical role in performance differentiation.

- **Topics:** Topic influence persists but lacks statistically significant patterns to identify dominant themes.

- **Aggregation methodology:** Sentiment aggregation methodology explain material performance variance, though optimal techniques vary by strategy.

These findings underscore the context-dependent nature of sentiment-driven portfolio construction while affirming the value of methodological rigor in signal extraction and aggregation.

6 Conclusion

This study investigates the integration of LLMs into financial sentiment analysis and portfolio construction, using a novel atom-based methodology. Summaries ('atoms') are generated from segmented earnings conference call transcripts of 69 software sector companies using GPT-3.5. Sentiment scores for these atoms are then computed using three approaches: DeBERTa within an NLI framework (at both atom and transcript levels), FinBERT, and a traditional lexicon-based method. Various sentiment aggregation techniques are explored, and both long-only and long-short portfolios are constructed using different weighting strategies informed by the sentiment scores.

This research makes three principal contributions to the literature. First, it demonstrates that the atom-based approach to sentiment extraction yields superior portfolio performance in long-only settings, underscoring the value of granularity in textual sentiment analysis. This result aligns with recent advances in financial NLP, suggesting that more detailed sentiment scoring can provide a more accurate assessment of underlying tone in financial disclosures. However, the methodology requires further refinement to consistently outperform in long-short frameworks. Second, this study finds that the choice of sentiment aggregation method has a substantial impact on portfolio outcomes, with the simple average (naive-average) approach outperforming more complex alternatives in most cases. Third, the results indicate that sentiment-weighted portfolio construction delivers higher returns and better risk-adjusted performance compared to traditional equal-weighted strategies.

Several avenues for future research emerge from these findings. The analysis could be extended by employing more recent or specialized LLMs, such as GPT-4, Mistral, or LLaMA 3.1, to assess whether improvements in language understanding translate into better sentiment extraction and portfolio performance. Expanding the dataset to include additional sectors or a broader universe of companies would enhance the generalizability of the results and allow for sector-based strategies, potentially incorporating market capitalization as an additional aggregation layer. Alternative portfolio construction techniques, such as market-cap weighting or risk-parity, could be explored to assess their interaction with sentiment signals. Further, topic modeling could be refined by clustering themes for more homogeneous groupings or by isolating highly sector-specific topics to evaluate their incremental value.

However, this study has several limitaions. The analysis is restricted to a sample of 69 software stocks, reflecting the constraints of the available dataset. This sector's historical outperformance may limit the generalizability of the results, particularly regarding the high absolute returns observed. Additionally, the use of GPT-3.5, while representative of current LLM capabilities, does not capture the latest advancements in language modeling, and future work should investigate whether newer models yield more robust sentiment signals. The inherent randomness of LLM outputs, even with deterministic settings, introduces some variability in sentiment scores and, consequently, in portfolio construction. Furthermore, the quarterly rebalancing frequency may not fully align with the timing of earnings releases, and future research could examine the impact of alternative rebalancing intervals or the lag between transcript publication and portfolio adjustment. Another consideration is the potential for look-ahead bias, as LLMs trained on data up to 2021 may inadvertently incorporate future information when evaluating historical transcripts. Techniques such as anonymizing product and company names could help assess the extent of this bias.

Finally, while this study focuses on earnings call transcripts, the methodology is readily extendable to other sources of financial text, such as MD&A sections, M&A conference call transcripts, or other regulatory filings. Prior research has demonstrated the informational value of these documents for investment decision-making (Feldman et al., 2008, Hu et al., 2021, Zhou et al., 2024). Applying the atom-based sentiment framework to a wider array of financial texts could further enhance the utility and robustness of sentiment-driven portfolio strategies.

In summary, this research advances the application of LLMs in financial sentiment analysis and portfolio management by demonstrating the benefits of granular sentiment extraction, careful aggregation, and innovative weighting schemes. While subject to certain data and methodological constraints, the findings provide a foundation for future work that leverages advances in language modeling and data availability to further improve sentiment-based investment strategies.

References

Araci, D., 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. Working paper, arXiv preprint, Cornell University, Ithaca, NY.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter Mood Predicts the Stock Market. Journal of Computational Science 2, 1–8.

Bommarito, M.J., Katz, D.M., 2022. GPT Takes the Bar Exam. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Curran Associates, Inc., Red Hook, NY, USA. pp. 1877–1901.

Burns, P.J., 2004. Performance Measurement Via Random Portfolios. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Chakraborty, I., Kim, M., Sudhir, K., 2019. Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes. Working paper, Cowles Foundation for Research in Economics, Yale University, New Haven, CT.

Chen, Y., Kelly, B.T., Xiu, D., 2022. Expected Returns and Large Language Models. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Cong, L., Liang, T., Yang, B., Zhang, X., 2019. Analyzing Textual Information at Scale. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? The Review of Financial Studies 22.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186.

Fama, E.F., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance 25, 383–417.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., Kyriazis, D., 2023. Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learning with Applications 14, 100508.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2008. The Incremental Information Content of Tone Change in Management Discussion and Analysis. SSRN Electronic Journal , 1–40.

Gentzkow, M., Kelly, B.T., Taddy, M., 2019. Text as Data. Journal of Economic Literature 57, 535–574.

Grefenstette, G., 1999. Tokenization. Springer Netherlands, Dordrecht. pp. 117–133.

Grossman, S.J., Stiglitz, J.E., 1980. On the Impossibility of Informationally Efficient Markets. The American Economic Review 70, 393–408.

He, P., Liu, X., Gao, J., Chen, W., 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, in: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).

Hu, M., Liu, B., 2004. Mining and Summarizing Customer Reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. ACM, New York, NY, USA. pp. 168–177.

Hu, W., Shohfi, T., Wang, R., 2021. What's Really in a Deal? Evidence from Textual Analysis of MA Conference Calls. Review of Financial Economics 39, 500–521.

Jha, M., Qian, J., Weber, M., Yang, B., 2024. ChatGPT and Corporate Policies. Working paper, Becker Friedman Institute for Economics, University of Chicago, Chicago, IL.

Kearney, C., Liu, S., 2014. Textual Sentiment in Finance: A Survey of Methods and Models. International Review of Financial Analysis 33, 171–185.

Kim, A., Muhn, M., Nikolaev, V., 2024. Bloated Disclosures: Can ChatGPT Help Investors Process Information? Working paper, Becker Friedman Institute for Economics, University of Chicago, Chicago, IL.

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification, in: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1746–1751.

Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A., 2009. Predicting Risk from Financial Reports with Regression, in: Ostendorf, M., Collins, M., Narayanan, S., Oard, D.W., Vanderwende, L. (Eds.), Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Boulder, Colorado. pp. 272–280.

Krimmel, M., Achterhold, J., Stueckler, J., 2024. Attention Normalization Impacts Cardinality Generalization in Slot Attention. Transactions on Machine Learning Research .

Lefort, B., Benhamou, E., Ohana, J.J., Saltiel, D., Guez, B., Jacquot, T., 2024. Sentiment Analysis of Bloomberg Markets Wrap Using ChatGPT: Application to the NASDAQ. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Li, F., 2006. Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? SSRN Electronic Journal .

Lopez-Lira, A., Tang, Y., 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2020. Textual analysis in finance. Annual Review of Financial Economics 12, 357–375.

Malo, P., Sinha, A., Takala, P., Korhonen, P., Wallenius, J., 2013. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. Working paper, arXiv preprint, Cornell University, Ithaca, NY.

Moskowitz, T.J., Ooi, Y.H., Pedersen, L.H., 2011. Time Series Momentum. Working paper, Chicago Booth School of Business; Fama-Miller Center; Becker Friedman Institute, Chicago, IL.

Paltoglou, G., Thelwall, M., 2010. A Study of Information Retrieval Weighting Schemes for Sentiment Analysis, in: Hajič, J., Carberry, S., Clark, S., Nivre, J. (Eds.), Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden. pp. 1386–1395.

Pelster, M., Val, J., 2024. Can ChatGPT assist in picking stocks? Finance Research Letters 59, 104786.

Phuong, M., Hutter, M., 2022. Formal Algorithms for Transformers. Working paper, arXiv preprint, Cornell University, Ithaca, NY.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training. Working paper, OpenAI, San Francisco, CA.

Schuettler, J., Audrino, F., Sigrist, F., 2024. Does sentiment help in asset pricing? A novel approach using large language models and market-based labels. Working paper, Social Science Research Network, Rochester, NY.

Tang, D., Qin, B., Liu, T., 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification, in: M‘arquez, L., Callison-Burch, C., Su, J. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal. pp. 1422–1432.

Tang, Y., Whitelaw, R.F., 2011. Time-Varying Sharpe Ratios and Market Timing. Quarterly Journal of Finance 1, 465–493.

Tetlock, P.C., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62, 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance 63, 1437–1467.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I., 2017. Attention Is All You Need, in: Advances in Neural Information Processing Systems 30 (NeurIPS 2017), pp. 5998–6008.

Wang, Y., Zhang, B., Zhu, X., 2018. The Momentum of News. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Yue, T., Au, D., Au, C.C., Iu, K.Y., 2023. Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology. Working paper, SSRN Electronic Journal, Social Science Research Network, Rochester, NY.

Zhou, W., Li, Y., Wang, D., Xueqin, D., Ke, Y., 2024. Management's tone in MDA disclosure and investment efficiency: Evidence from China. Finance Research Letters 58, 104767.

Appendix

	Mean Return	Median Return	Standard Deviation	Skewness	Kurtosis
Minimum	-0.131	-0.117	0.098	-0.802	-1.890
25^{th} Quantile	0.037	0.013	0.152	-0.140	-0.508
Mean	0.057	0.044	0.240	0.374	0.824
Median	0.057	0.059	0.207	0.213	0.115
75^{th} Quantile	0.082	0.083	0.303	0.799	1.153
Maximum	0.162	0.175	0.506	2.249	9.267

Table 14: Descriptive Statistics of Investment Universe Returns (Quarterly)
This table presents key distributional characteristics for quarterly returns of the 69-stock software sector investment universe, covering the sample period from Q1 2013 to Q4 2023. Reported metrics include mean return, median return, standard deviation (volatility), skewness (asymmetry measure), and kurtosis (tail extremity measure).

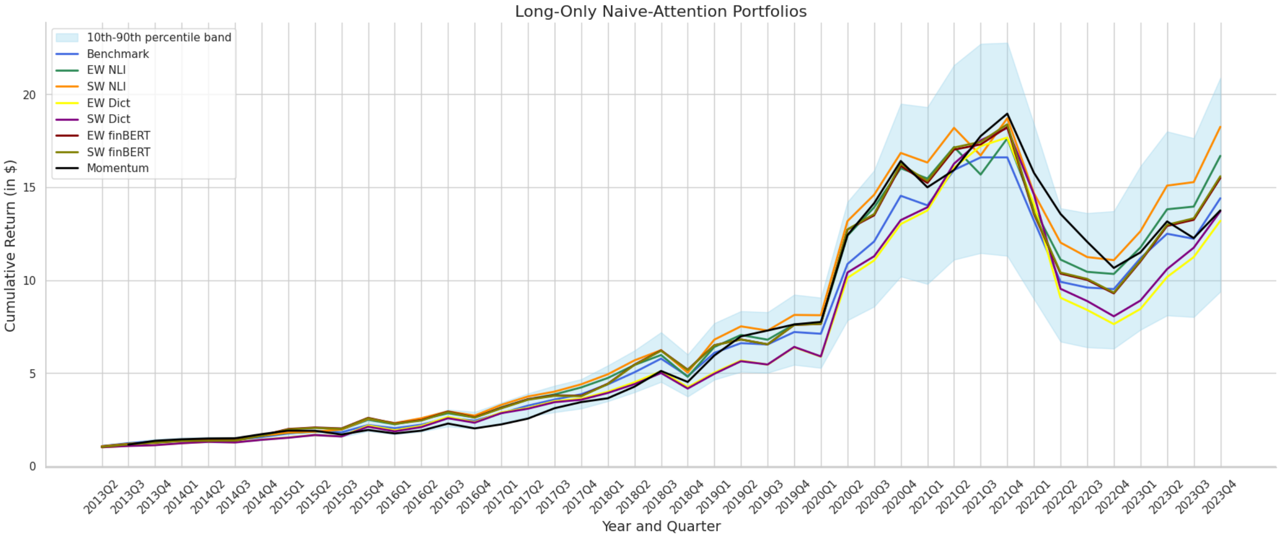

Figure 2: Cumulative Returns of Long-Only Naive-Attention Portfolios This figure displays the cumulative returns of long-only portfolios constructed using the naive-attention sentiment aggregation method, evaluated with both EW and SW schemes across the three sentiment extraction approaches: NLI (NLI-t), FinBERT (FB), and Dictionary-based methods (MD). The results are compared to a benchmark, a momentum strategy, and the 10th–90th percentile band of randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis reports cumulative return in dollars.

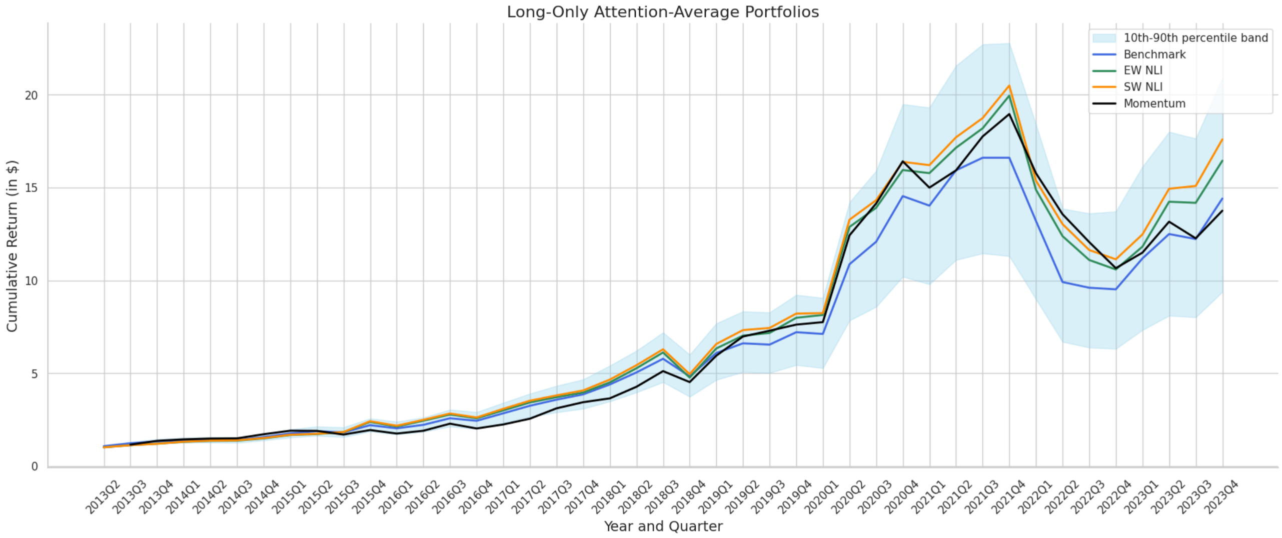

Figure 3: Cumulative Returns of Long-Only Attention-Average Portfolios
This figure presents the cumulative returns of long-only portfolios constructed using the attention-average sentiment aggregation method. Results are shown for both EW and SW strategies based on NLI (NLI-t) sentiment extraction, alongside the benchmark and a momentum strategy. The shaded area represents the 10th–90th percentile band of cumulative returns from randomly generated portfolios. The x-axis indicates the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

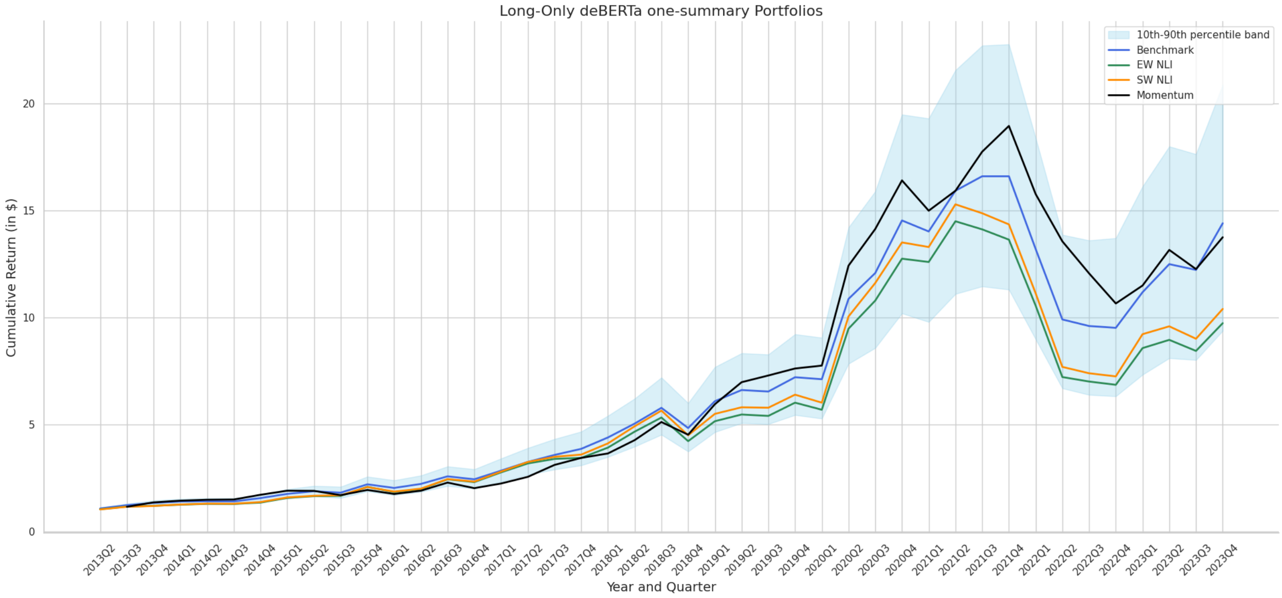

Figure 4: Cumulative Returns of Long-Only One-Summary Portfolios
This figure displays the cumulative returns of long-only portfolios constructed using DeBERTa-based sentiment scores extracted from one-summary representations of earnings call transcripts. Results are shown for both EW and SW strategies, alongside the benchmark and a momentum strategy. The shaded area represents the 10th–90th percentile band of cumulative returns from randomly generated portfolios. The x-axis indicates the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

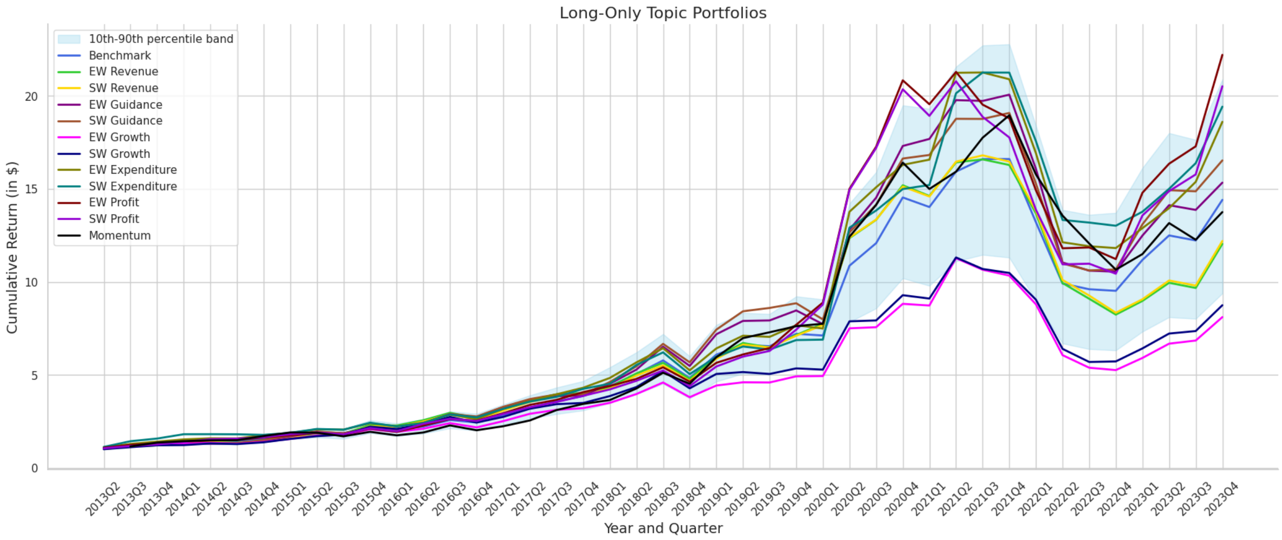

Figure 5: Cumulative Returns of Long-Only Topic Portfolios This figure presents the cumulative returns of long-only portfolios constructed based on sentiment scores for five key topics: Revenue, Guidance, Growth, Expenditure, and Profitability. Results are shown for both EW and SW strategies for each topic. The performance of these portfolios is compared to a benchmark, a momentum strategy, and the 10th–90th percentile band of randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

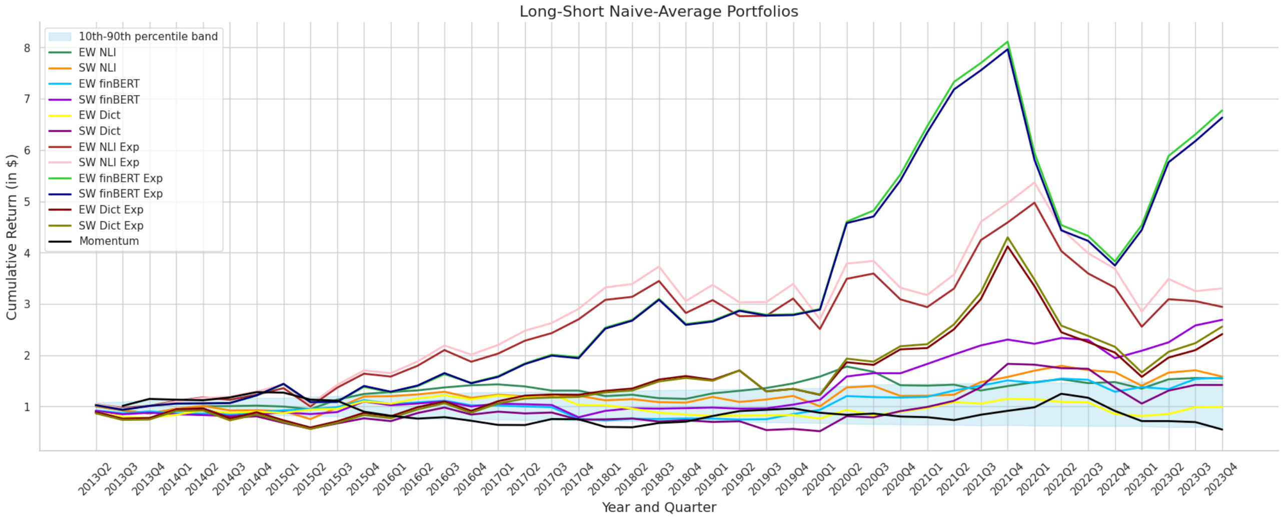

Figure 6: Cumulative Returns of Long-Short Naive-Average Portfolios This figure displays the cumulative returns of long-short portfolios constructed using the naive-average sentiment aggregation method. Results are shown for both EW and SW strategies across the three sentiment extraction approaches: NLI (NLI-t), FinBERT (FB), and Dictionary (MD), as well as their expanded (Exp) variants. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis reports cumulative return in dollars.

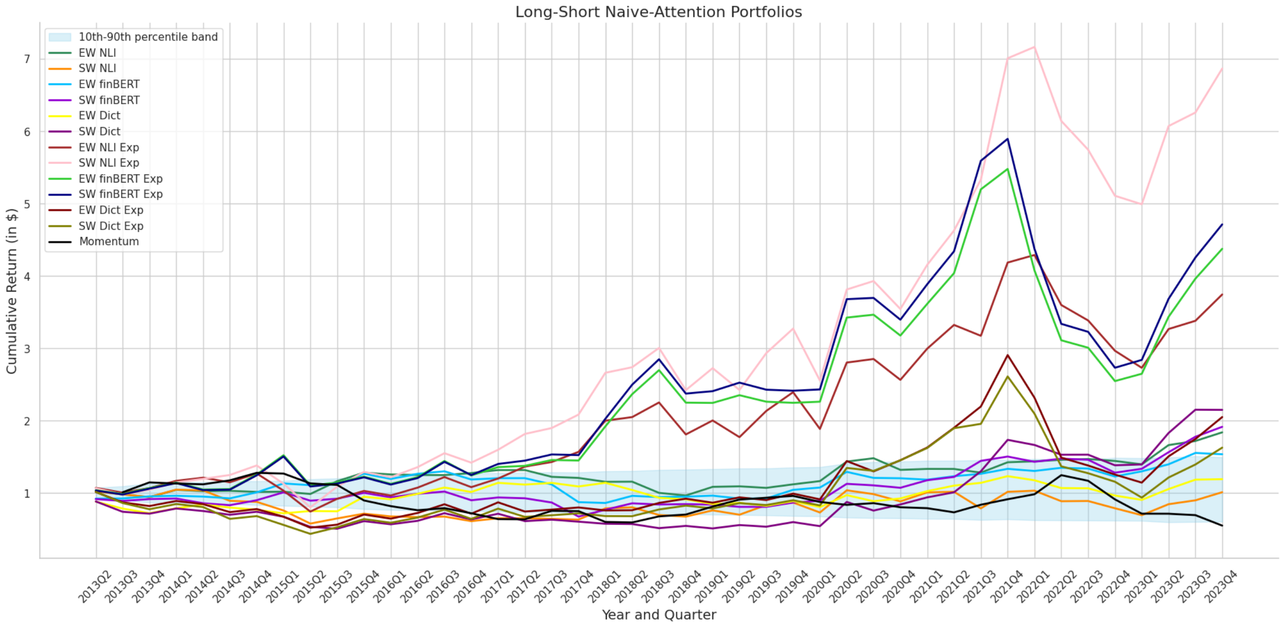

Figure 7: Cumulative Returns of Long-Short Naive-Attention Portfolios This figure displays the cumulative returns of long-short portfolios constructed using the naive-attention sentiment aggregation method. Results are shown for both EW and SW strategies across three sentiment extraction approaches: NLI (NLI-t), FinBERT (FB), and Dictionary (MD), as well as their expanded (Exp) variants. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

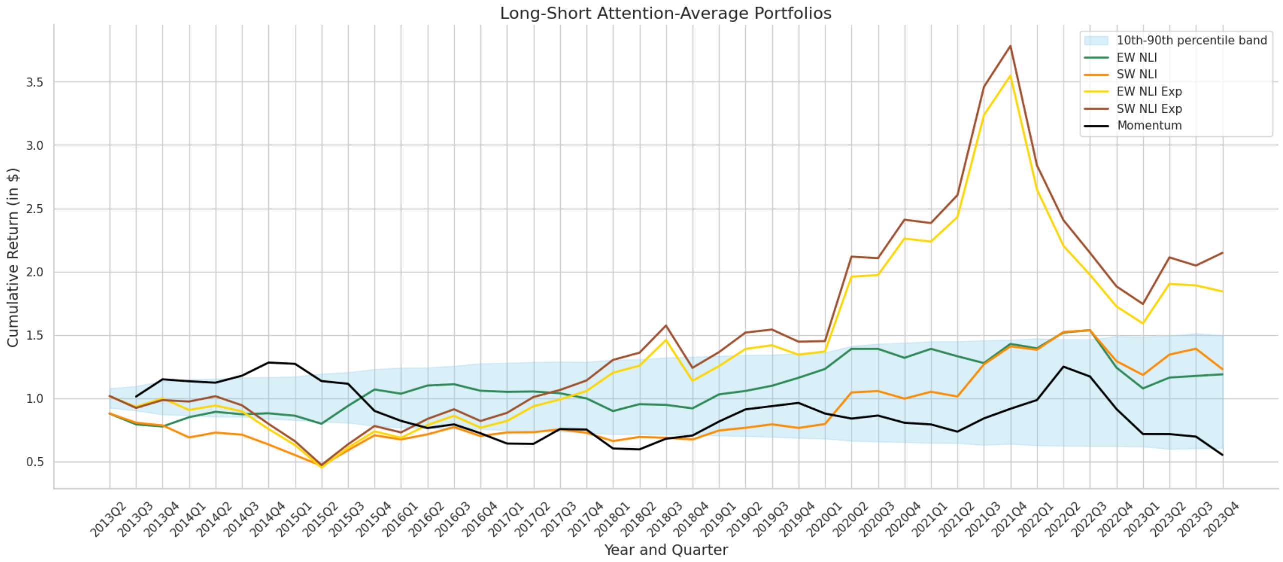

Figure 8: Cumulative Returns of Long-Short Attention-Average Portfolios This figure displays the cumulative returns of long-short portfolios constructed using the attention-average sentiment aggregation method. Results are shown for both EW and SW strategies across three sentiment extraction approaches: NLI (NLI-t), FinBERT (FB), and Dictionary (MD), as well as their expanded (Exp) variants. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

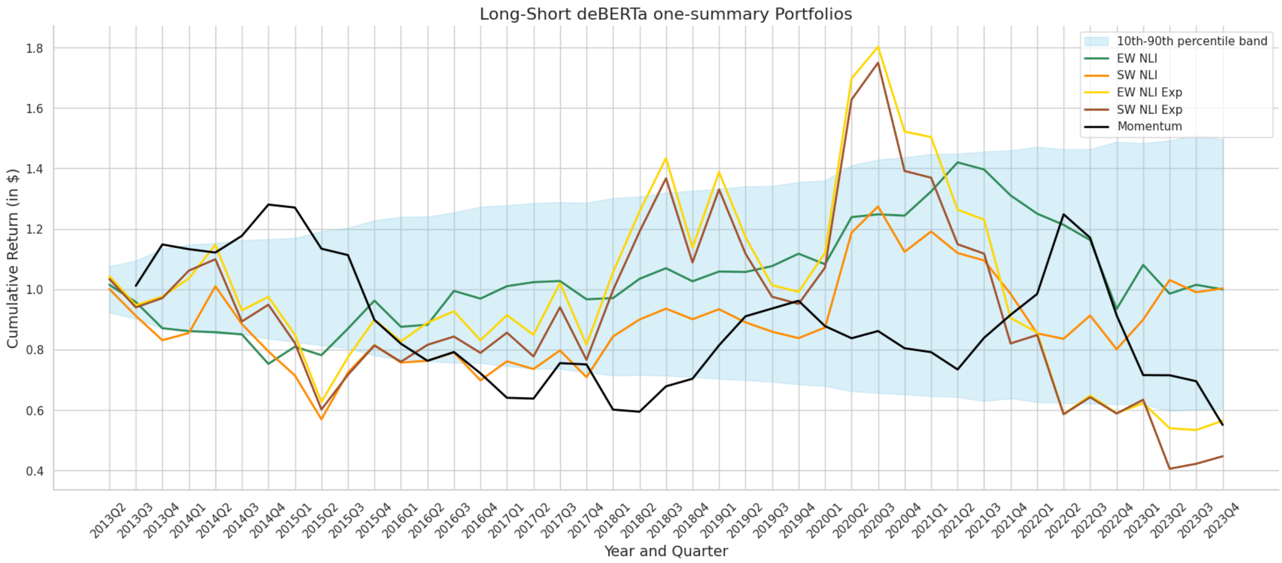

Figure 9: Cumulative Returns of Long-Short One-Summary Portfolios This figure displays the cumulative returns of long-short portfolios constructed using DeBERTa-based sentiment scores derived from one-summary representations of earnings call transcripts. Results are shown for both EW and SW strategies, as well as their expanded (Exp) variants. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

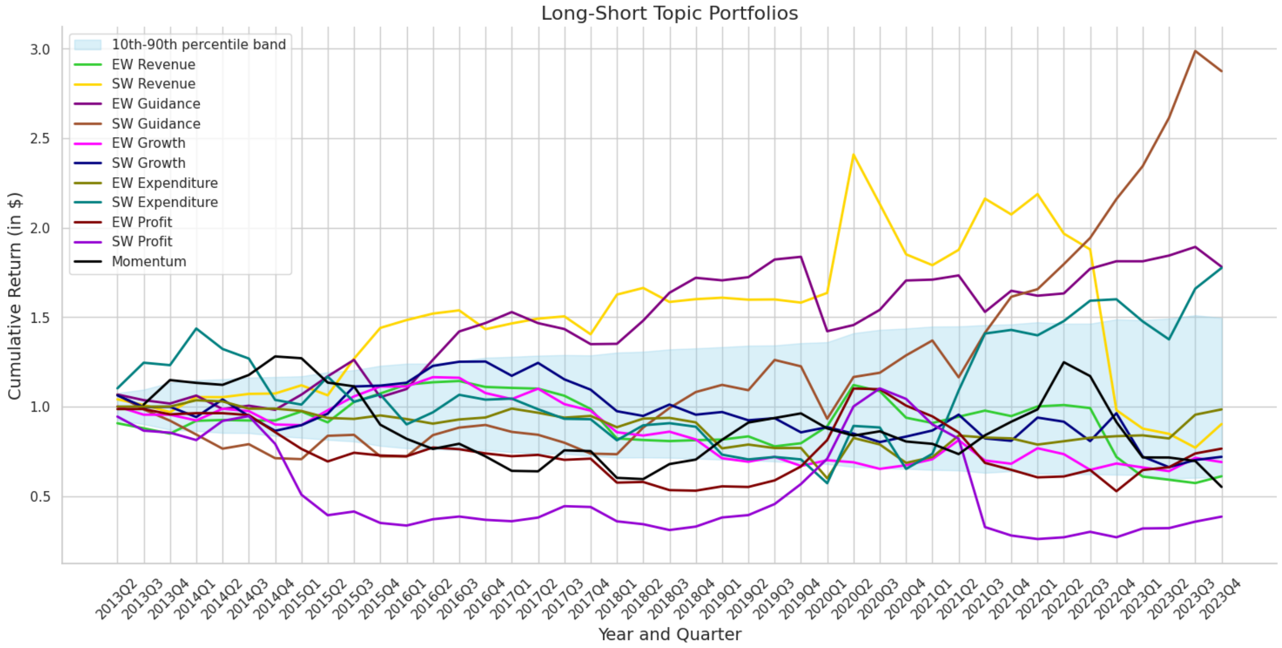

Figure 10: Cumulative Returns of Regular Long-Short Topic Portfolios This figure presents the cumulative returns of the regular long-short portfolios constructed based on sentiment scores for five key topics: Revenue, Guidance, Growth, Expenditure, and Profitability. Results are shown for both EW and SW strategies for each topic. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.

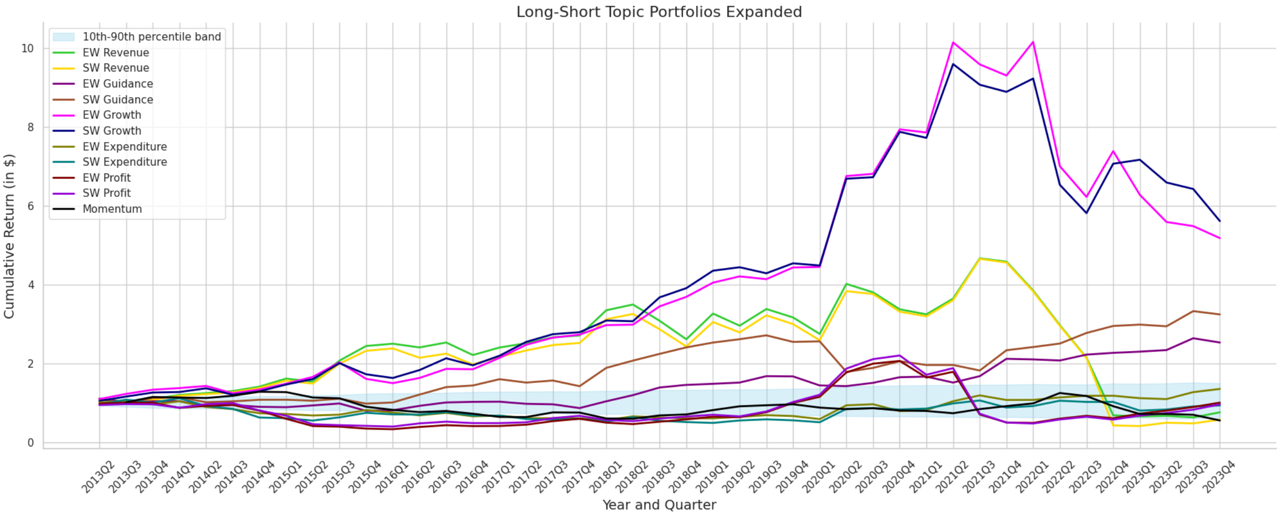

Figure 11: Cumulative Returns of Expanded Long-Short Topic Portfolios This figure presents the cumulative returns of the expanded long-short portfolios constructed based on sentiment scores for five key topics: Revenue, Guidance, Growth, Expenditure, and Profitability. Results are shown for both EW and SW strategies for each topic. The performance of these portfolios is compared to a momentum strategy and the 10th–90th percentile band of cumulative returns from randomly generated portfolios (shaded area). The x-axis represents the evaluation period by quarter, and the y-axis displays cumulative return in dollars.