Cureus Journal of Business and Economics

Open Access Original Article

Linguistic Alphas: Decoding the Market Impact of Words in Software Earnings Calls

Jose Juan De Leon [1] , Francesca Medda [1]

1. Institute of Finance and Technology, University College London, London, GBR

Corresponding author: Jose Juan De Leon, jose.guillamon.19@ucl.ac.uk

Received 06/29/2025
Review began 07/06/2025
Review ended 09/28/2025
Published 11/13/2025

DOI:
https://doi.org/10.7759/s44404-025-08244-6

Abstract

Markets do not just move on numbers, they move on words. A single phrase in an earnings call can shift billions in market value, yet traditional financial models overlook the power of language. This study uncovers a hidden financial signal in corporate speech, analyzing 437 Software as a Service companies and 40,574 unique words from earnings calls (2003-2024).

We introduce a context-aware linguistic model that categorizes words into technicisms, proper names, and mundane words, capturing how subtle phrasing shifts investor sentiment. By distinguishing between commonly overlooked linguistic structures and industry-specific jargon, our approach refines the interpretation of financial discourse. A novel weighting methodology enhances sentiment classification by 74% (experimental) and 42% (theoretical) over traditional models, improving market prediction beyond the Fama French 5 factor model. The impact is substantial. A portfolio built on these linguistic insights yields 30% annualized returns versus the market's 20%, with a Beta of 0.78. These findings prove that language is not just a reflection of market sentiment, it is an investable asset class. Our results demonstrate that even seemingly neutral words influence investor perception, affecting stock performance in ways conventional models fail to capture.

By integrating financial linguistics with asset pricing, this study provides a scalable and adaptable framework for extracting predictive signals from textual data. As Artificial Intelligence-driven financial analysis becomes more sophisticated, mastering the language of finance will be essential for maintaining a competitive edge in market prediction.

Categories: Computational Intelligence and Information Management, Banking and financial services, Financial reporting
Keywords: experimental finance, comparable analysis, comparable valuation, equity analysis, equity valuation, factor-based models, value investing

Introduction

When Jerome Powell (chair of the U.S. Federal Reserve) speaks, the market does not just hear, it deciphers. Our research focuses on earnings calls, where finance emerges as a latent language where meaning is derived not only from individual words but also from their combinations and context. A single phrase may shift market sentiment due to the underlying financial significance it encodes, underscoring the precision required to interpret such communications (Tetlock et al., 2008).

The research gaps and contributions of this paper are as follows. Prior studies demonstrate that tone and generic sentiment influence markets (Tetlock et al., 2008) (Loughran and McDonald, 2011), but they underweight two features central to software earnings calls: (i) domain-specific technicisms and (ii) context-sensitive phrasing. Words that appear neutral or negative in isolation can shift meaning in context, while sector-specific jargon conveys information that generic lexicons often miss.

To address these gaps, we introduce a context-aware, industry-specific linguistic model that classifies terms into Technicisms, Proper Names, and Mundane words and assigns sentence-level sentiment. Beyond classification, we link these linguistic signals to traded factors, demonstrating their explanatory and predictive value in financial markets.

Positioning within the literature, our study bridges three strands: lexicon methods that lack context sensitivity, transformer models that capture context but are rarely linked to traded factor tests, and sector-agnostic studies with short samples. By preserving industry vocabulary, assigning sentence-level sentiment, and testing against Fama-French factors on a long sample, we provide evidence that domain language and context add information beyond tone.

The complexity of financial discourse is heightened by its inherent ambiguity and context dependence. For example, while standard financial dictionaries offer clear definitions for terms like CAPEX (Capital

Expenditure), they often fall short when it comes to industry-specific expressions. In sectors such as technology and SaaS (Software as a Service), terms like OEM (Original Equipment Manufacturer) lack comprehensive definitions, complicating the interpretation of domain-specific language. This challenge is further compounded by the limitations of conventional natural language processing techniques. Methods that expand abbreviations risk fragmenting multi-word technicisms, while bigram-based models may oversimplify complex phrases, potentially misclassifying terms such as “decline.”

Consider the following example:

"Our OEM production experienced a sharp decline in profitability this year due to elevated CAPEX pressures, but the subsequent reduction in OPEX is expected to significantly lower costs, decrease operational inefficiencies, and drive substantial long-term revenue expansion."

In this instance, while CAPEX is widely recognized, OEM remains ambiguous, highlighting the multifaceted challenges in deciphering financial communications. Traditional sentiment analysis methods often assign static sentiment values to words without accounting for contextual nuances (Loughran and McDonald, 2011). As a result, terms such as “decline” may be misclassified, since the same word can represent negative outcomes in one context and positive cost-saving measures in another.

Furthermore, although bigram and n-gram models, as discussed in standard NLP literature (Garcia et al., 2023), strive to preserve contextual integrity by analyzing word pairs, they can inadvertently fragment key expressions and fail to capture the interplay of multi-word phrases. Even advanced machine learning models such as BERT and GPT-3, which have greatly enhanced our understanding of semantic structures (Devlin et al., 2019), (Brown et al., 2020), (Buehlmaier and Whited, 2018), and (Loughran and McDonald, 2016), may lack the granularity required to fully interpret the layered nuances of domain-specific language.

To address these challenges, we propose an integrated approach that combines advanced machine learning techniques with bespoke financial lexicons. Our methodology incorporates a suite of specialized dictionaries, the Technicism Dictionary to capture sector-specific jargon, the Mundane Words Dictionary to recognize sentiment-modifying terms, and the Proper Nouns Dictionary to account for influential figures. This framework allows for a context-aware interpretation of sentiment, where the same word may convey different implications depending on its usage. By shifting the focus from solely capturing managerial tone to examining thematic words that directly influence market behavior, our approach not only refines sentiment classification but also offers actionable insights into investor reactions and market performance.

In doing so, it integrates modern machine learning techniques with finance-specific linguistic precision, offering both enhanced theoretical insight and practical applicability. It comprehensively addresses the shortcomings of existing models by incorporating both industry-specific terminology and previously overlooked linguistic modifiers. Our work leverages a uniquely extensive 21-year dataset (Davis and Tama-Sweet, 2011), despite significant challenges such as size limitations that required stringent filtering, the computational intensity of processing large volumes of transcripts, and variability in transcript quality over time.

Prior research has established that linguistic clarity and tone affect market outcomes. Opaque or complex disclosures are associated with higher uncertainty and weaker firm performance (Huang et al., 2014), (Feng, 2010), (Price et al., 2012), and (Feldman et al., 2009). Seminal studies demonstrate that tone and generic sentiment influence asset prices, with media and textual signals providing predictive content for returns (Tetlock et al., 2008). Moreover, trading strategies that exploit sentiment-driven signals have been shown to generate excess returns, highlighting the practical relevance of language in shaping investor behavior (Jegadeesh and Wu, 2013). Collectively, this body of research confirms the role of linguistic factors in financial markets but primarily relies on general-purpose sentiment dictionaries, shorter time horizons, or sector-agnostic samples.

Our study extends this literature by explicitly addressing these limitations. Whereas existing work often treats words as static carriers of sentiment, we demonstrate that contextual phrasing and domain-specific technicisms alter interpretation, particularly in technology and SaaS earnings calls. By preserving sector vocabulary, assigning sentence-level sentiment, and validating these signals against Fama-French factors, the established asset pricing benchmarks, we provide evidence that industry-specific language carries additional information beyond tone. Furthermore, by applying this framework to an unusually extensive 21-year dataset of U.S. SaaS earnings calls, our analysis shows that the informational content of financial discourse can be systematically linked to traded factors over long horizons. In this way, the paper bridges prior lexicon- and sentiment-based approaches with modern context-aware methods, offering both theoretical refinement and practical trading implications.

Our paper proceeds as follows: The Research method section starts by detailing the dataset used in the

study and the preparation process for analysis to be continued by the outline of the methodology employed. The Results and Discussion section presents the results and the experimental validation of our approach, incorporating the findings to develop a trading strategy based on sentiment-driven factors. Finally, Conclusions summarize the paper by giving key insights and their implications.

Research Method

Data and data preparation

We start our research with data gathering and preparation. In this section, we first look at the data gathered and its descriptive analytics; we then move to observing how the data are prepared to be employed in the analysis. The data preparation stage is the most critical stage, being Transforming Unstructured Data into Structured Formats due to the unstructured nature of the data examined. In this section, we first examine the data employed and then examine its preparation.

The data employed for our analysis consist of earnings call reports obtained from the population of North American traded equities, including stocks traded within the NYSE and NASDAQ. The analyzed data ranges from February 2003 to December 2023. The data are obtained from Refinitiv Eikon by using the Event Screener function, an application that allows the identification of companies with the specified criteria, and includes all publicly traded SaaS firms within the period of analysis (Refinitiv Eikon, 2024). If companies were at any point exchange-traded and now are traded Over the Counter or delisted, these are also included.

The data are obtained in an unstructured format as the Earnings Calls transcript from Refinitiv Eikon and employed natural language processing to transform these data into a structured format (Refinitiv Eikon, 2024). The earnings call transcripts were obtained in an unstructured text format, necessitating transformation into a structured dataset suitable for financial analysis. This transformation was performed using Natural Language Processing (NLP) techniques; this is later covered in the methodology section. Both daily and monthly market returns employed for the short- and long-term impact analysis were obtained from the CRSP Stock/Security Files within Wharton Research Data Services (Wharton Research Data Services, 2024).

To address survivorship bias in our analysis, we review the inclusion of equities on a monthly basis. This approach ensures that our dataset only encompasses those companies that are publicly traded in the specific month under analysis. This methodology reduces the impact of survivorship bias, which can skew analysis results by only accounting for companies that have survived through to the end of the period under study. While some level of bias may still be present within the confines of a single month, the effect is significantly minimized compared to evaluations based on longer periods such as yearly or quarterly assessments.

Furthermore, all choices and allocations within our dataset are assessed on a monthly basis, starting retrospectively from the earliest available data. This method ensures that each decision or selection is grounded in the historical context of the specific month, avoiding any forward-looking biases and providing a more accurate reflection of the market conditions at that time. It is important to note that if daily returns were used for this analysis, the survivorship bias would be confined to just the month, as the presence of the equity in the principal indexes is only checked monthly. The statistics relating to the research corpus are presented in Table *1*.

Earnings calls	Value
Start Date	03.02.2003
End Date	21.12.2023
Unique Firms	437
Observations	9962
Average Words per Document	7898

TABLE 1: Research corpus overview. The start and end dates indicate the period of analysis, unique firms represent the number of different firms analyzed across the research, observations refer to the total number of earnings calls analyzed, and average words per document is the mean number of words across all earnings calls

We now delve into the data preparation process, consisting of several key stages. First, text preprocessing ensures that unstructured data is cleaned and refined for analysis. Next, segmentation and keyword identification organize the text by grouping related content based on thematic relevance.

Text Preprocessing

The initial phase of our methodology involves cleaning the textual data by removing noise, such as punctuation and special symbols, and standardizing all text to lowercase. This normalization is crucial for ensuring consistency in subsequent analyses. Sun et al. (Sun et al., 2014) highlight that effective preprocessing, encompassing noise reduction and structured text organization, enhances the usability of financial texts for feature extraction and sentiment analysis.

In line with Sun et al. (Sun et al., 2014) and Haddi et al. (Haddi et al., 2013), we preprocess the data by removing punctuation and special symbols while converting text to lowercase to maintain consistency. However, we do not strictly adhere to all preprocessing recommendations found in the literature. Our approach incorporates an advanced NLP system with sentence transformers, enabling us to interpret numerical values-an aspect that traditional methods often recommend removing. Eliminating numbers indiscriminately could lead to a loss of critical information; for example, removing "50%" from the phrase "our sales have grown by 50%" would hinder our model's ability to accurately capture financial insights.

Furthermore, existing preprocessing literature primarily focuses on written financial disclosures rather than spoken ones. Conventional recommendations, such as removing proper names, may not be suitable for our analysis, as names can carry important contextual meaning in spoken disclosures. Given these distinctions, we tailor our preprocessing steps to better align with the nature of spoken financial communication.

Additionally, analyzing word frequency across documents and retaining only those that appear in a significant portion of earnings calls (e.g., at least 5%) helps focus on terms with substantial relevance. This approach reduces the impact of outliers and rare terms, aligning with methodologies that emphasize prevalent linguistic features for more robust predictive modeling.

To implement this in our study, we analyzed and counted all words across earnings reports, retaining only those that appeared in at least 5% of earnings calls. This ensures that our analysis captures meaningful patterns while filtering out less relevant terms.

Segmentation and Keyword Identification

The segmentation of earnings call transcripts into individual sentences is essential for contextual analysis, allowing for the extraction of word-specific sentiment measures at a granular level. Sentence segmentation has been widely adopted in financial text processing, as aggregating sentiment over entire documents may dilute the impact of localized information, thereby obscuring the relationship between textual features and asset prices (Todd et al., 2024), (Brockman et al., 2015), and (Price et al., 2012). Studies show that sentence-based segmentation improves the explanatory power of textual sentiment models by preserving context while reducing noise from non-informative sections (Todd et al., 2024) (Fu et al., 2021).

Todd et al. (Todd et al., 2024) emphasize that granular sentiment extraction, particularly in earnings calls, enhances the ability to detect managerial tone shifts that impact stock prices. Likewise, Fu et al. (Fu et al., 2021) demonstrate that sentence-level sentiment measures in earnings calls can predict stock price crash risk, reinforcing the importance of refined segmentation techniques. Following the previously cited research, the cleaned data were segmented into sentences, facilitating targeted analysis.

Additionally, a key contribution of this study is the categorization of words into three distinct linguistic groups, each offering unique insight into the structure and sentiment of earnings calls. The first category, mundane words, includes frequently used but contextually relevant terms that provide structural coherence without conveying explicit sentiment. While often overlooked, these words shape the tone and delivery of managerial communication, subtly influencing investor interpretation (Loughran and McDonald, 2011) (Price et al., 2012). The second group, proper names, encompasses references to executives and key stakeholders, enabling sentiment analysis tied to specific individuals. Research shows that investor reactions are significantly affected by the perceived tone and demeanor of these figures during earnings calls (Mayew and Venkatachalam, 2012) (Brockman et al., 2015). Lastly, the technicism category captures domain-specific terminology pertinent to SaaS firms and financial markets, ensuring that constructed factors remain anchored in industry-relevant discourse (Loughran and McDonald, 2011) (Henry, 2006).

To illustrate, Table *2* presents examples drawn directly from earnings call transcripts, showing how phrases can be decomposed into mundane terms ("going," "lower"), proper names ("Jim"), or technicisms

("traffic," "retention," "growth"). These cases highlight why classification matters: mundane words can frame tone without explicit sentiment, proper names anchor responsibility to specific executives (e.g., "Thanks, Jim"), and technicisms carry sector-relevant financial implications such as revenue growth or client retention. By separating these categories, the algorithm reduces noise from high-frequency but low-information terms, distinguishes personal attributions from general commentary, and ensures that industry-specific signals are preserved rather than diluted by generic lexicons. This layered structure improves interpretability and increases the likelihood that extracted sentiment factors reflect information investors actually process.

Company	Quarter	Sentence fragment (from transcript)	Mundane words (from Mundane Dictionary)	Proper names (from Proper Nouns Dictionary)	Technicisms (from Technicism Dictionary)	Context identification
Akamai Technologies	Q1 2024	Industry traffic is lower than we had initially expected.	lower; than; expected; had		traffic	Negative
Akamai Technologies	Q1 2024	Their observability solution enables … real-time data ingestion … and extensive data retention …			retention	Positive
Adobe Inc.	Q1 2024	Subscription revenue was $1.16 billion, representing 12% year-over-year growth.	was; growth			Positive
Broadridge Financial Solutions, Inc.	Q1 2019	Our business continues to benefit from new client additions as well as higher trading revenue and a growth of professional services revenues.	continues; higher		growth	Positive
Broadridge Financial Solutions, Inc.	Q1 2019	And as Tim pointed out, we're going to do that all at a lower cost still.	going; out; do; lower; cost; still			Positive
Broadridge Financial Solutions, Inc.	Q1 2019	Thanks, Jim, and good morning, everyone.	good	Jim		Positive

TABLE 2: Example classification by company and quarter

This classification framework enhances the information content of text-based factors by distinguishing common narrative elements from domain-specific linguistic signals, a methodology shown to improve the predictive power of textual data in financial applications (Todd et al., 2024), (Bochkay et al., 2020), and (Chen et al., 2018). Prior research has demonstrated that extreme language in earnings calls is associated with increased market volatility (Bochkay et al., 2020), while differences in sentiment expressed by managers and analysts convey distinct informational signals to investors (Chen et al., 2018). Collectively, these findings suggest that refining textual analysis through careful segmentation and categorization significantly enhances the ability to anticipate market movements.

Method

The proposed methodology consists of several steps that can be organized into distinct sections, presented in the order of execution. These sections include: Sentiment Identification, Individual Word Portfolio Fit Assessment, Equities Scores Attribution, and Weighted Factor Construction. Each step builds upon the previous one to ensure a comprehensive and structured approach to the analysis. The data are split into two datasets: 2003 to 2019 forming part of the training dataset, and 2019 to 2023 forming part of the test dataset. All analyses below will be performed in the training dataset, unless stated otherwise. The following methodology is employed in all of the word baskets identified independently.

Sentiment Identification

For sentiment analysis, the distilbert-base-uncased-finetuned-sst-2-english model was used. This model, which is fine-tuned on the Stanford Sentiment Treebank (SST-2), is effective for classifying short texts into "positive" or "negative" sentiment categories with a high degree of accuracy (Todd et al., 2024), (Sun et al., 2014), and (Devlin et al., 2019). Prior research has shown that transformer-based models such as

DistilBERT and BERT have contributed significantly to improvements in financial sentiment analysis (Todd et al., 2024). Studies have demonstrated that BERT-based architectures achieve strong performance across a wide range of NLP tasks, making them suitable for sentiment classification (Devlin et al., 2019). Additionally, fine-tuned transformer models have been shown to outperform traditional lexicon-based approaches in detecting subtle differences in sentiment (Sun et al., 2014).

Sentiment analysis was applied to keyword-specific contexts, allowing for a quantitative measure of sentiment polarity associated with the language used in earnings calls (Chen et al., 2018), (Bochkay et al., 2020), (Brockman et al., 2015), (Kearney and Liu, 2014), and (Loughran and McDonald, 2011). Prior studies show that sentiment derived from manager and analyst interactions in earnings calls can predict intraday stock price movements (Sun et al., 2014). Sentiment intensity, particularly when expressed through extreme language, has also been found to influence trading volumes (Bochkay et al., 2020). Outside the context of earnings calls, a comprehensive review of sentiment analysis methods in finance demonstrates how sentiment models influence asset prices across a variety of financial text sources, including news articles and analyst reports (Kearney and Liu, 2014). Furthermore, researchers have emphasized the importance of domain-specific sentiment dictionaries. General-purpose sentiment lexicons often misclassify financial terms, creating challenges when applying standard sentiment models in financial analysis (Loughran and McDonald, 2011).

The processed results were aggregated at the company and keyword level, producing structured metrics such as:

-Total keyword occurrences.
-Sentiment counts (positive and negative).
-Average sentiment scores.

The structured dataset thus derived enables quantitative and thematic analyses, linking qualitative insights from unstructured data to measurable financial outcomes (Todd et al., 2024), (Kearney and Liu, 2014), and (Loughran and McDonald, 2011). By leveraging an NLP model, this methodology ensures a rigorous and replicable approach to analyzing financial text (Jegadeesh and Wu, 2013), (Sun et al., 2019), (Bochkay et al., 2020), and (Chen et al., 2018).

Portfolio Compilation

Prior research suggests that linguistic complexity in corporate disclosures can stem from two competing forces: informative disclosure or obfuscation (Feng, 2010) (Bushee et al., 2017). In earnings calls, verbosity can indicate either higher informational content or efforts to mask poor performance through complexity (Loughran and McDonald, 2014). To systematically assess how variation in linguistic complexity relates to stock performance, equities were categorized into three groups based on the volume of words spoken during earnings calls.

At the end of each month, equities were reclassified based on the rolling distribution of word count in earnings calls over the prior period. The first category consisted of equities in the bottom 33% of words spoken, representing firms with lower disclosure intensity. The second category captured the middle 33%, while the third included the top 33%, corresponding to firms with high disclosure and potentially more complex financial narratives. This classification methodology follows empirical research suggesting that textual attributes-such as sentence complexity, word count, and verbosity-can systematically influence investor interpretation and subsequent market reactions (Feng, 2010) (Bushee et al., 2017).

Additionally, a dynamic reclassification method was used, ensuring that each equity's categorization remains reflective of the most recent earnings call data, preventing outdated classification effects. This rolling approach aligns with portfolio sorting techniques in empirical asset pricing research (Fama and French, 1993) (Jegadeesh and Titman, 1993) and ensures that linguistic factors are linked to contemporaneous market data rather than static historical classifications.

Individual Word Portfolio Fit Assessment

The raw text data (individual word counts from earnings reports) is first cleaned and aligned with the Fama-French factors and portfolio returns, ensuring all datasets are grouped by month. This follows the previously presented research from the first stage of the methodology.

Individual word occurrences are grouped into portfolios based on quartile rankings, with words exhibiting zero frequency classified separately as quartile 0. Each quartile represents a distinct category of words ranked by their frequency, ensuring that the distribution of words across quartiles reflects varying levels of textual significance. The quartiles are constructed such that each contains an equal or approximately equal number of word occurrences, allowing for a balanced segmentation that facilitates meaningful comparative analysis. This methodology is aligned with prior research in financial textual analysis, where

structuring text-based variables into quantifiable groups improves the interpretability and predictive power of sentiment models (Loughran and McDonald, 2011) (Loughran and McDonald, 2014). The ranking of words based on their occurrence follows similar approaches in financial sentiment analysis, where words are categorized by frequency to assess their impact on asset pricing and market behavior (Tetlock, 2007) (Garcia, 2013).

Factors that fall entirely within a single quartile are excluded from further analysis, as their lack of variation provides insufficient statistical power for robust inference. The exclusion of such factors aligns with best practices in empirical finance, where variables with minimal dispersion fail to offer meaningful explanatory insights (Bloomfield, 2008) (Biddle et al., 2009). Similar to factor-based investing, where low-variance factors are often removed to prevent statistical distortions (Fama and French, 1993), our approach ensures that only variables with significant variation across quartiles are retained for further analysis. By filtering out these non-informative factors, the methodology strengthens the reliability of textual analysis in financial modeling, allowing for a more structured evaluation of how varying levels of word frequency contribute to market sentiment and investment decisions.

The Gibbons-Ross-Shanken (GRS) test is applied to evaluate the joint significance of individual word portfolios, focusing on their explanatory power when combined with the Fama-French five-factor model (Fama and French, 2015). This test provides a rigorous statistical framework for assessing whether the inclusion of word-based factors improves the model's ability to explain portfolio returns beyond traditional risk factors.

To assess the explanatory power of specific quartiles, each previously computed factor is regressed alongside the five Fama-French factors-market (MKT), size (SMB), value (HML), profitability (RMW), and investment (CMA)-as specified in the equation below. The additional factor is sequentially exchanged and tested against all previously generated factors to determine its relative contribution. This step ensures that only the most statistically relevant factors are retained, aligning with the framework of evaluating whether a set of factors jointly explains asset returns in a mean-variance efficient manner (Gibbons et al., 1989).

The model equation is presented as follows (equation 1):

$$\begin{aligned} E(R_i) - R_f = \alpha_i + \beta_i \cdot (E(R_m) - R_f) + \beta_j \cdot \mathrm{SMB}_t + \beta_k \cdot \mathrm{HML}_t \\ + \beta_l \cdot \mathrm{RMW}_t + \beta_m \cdot \mathrm{CMA}_t \\ + \beta_s \cdot (\mathrm{ChangingTerm}_t - (E(R_m) - R_f)) + e_{it} \end{aligned} \tag{1}$$

The variables are described as follows: $E(R_i) - R_f$ represents the excess returns of the previously computed population portfolios over the risk-free rate, capturing the additional gain expected from the portfolio compared to a secure investment. $E(R_m) - R_f$ designates the excess market returns, capturing the overall market's performance above the risk-free rate. SMB_t represents the excess returns of small-capitalization companies over large-capitalization companies, highlighting the size effect in asset pricing. HML_t quantifies the excess returns of stocks with high book-to-price ratios over those with low book-to-price ratios, indicating the value premium. RMW_t measures the difference in returns between diversified portfolios of stocks with robust and weak profitability. CMA captures the differential returns between portfolios of low and high investment firms. $ChangingTerm_t$ is a placeholder for the introduced factors. e_{it} accounts for idiosyncratic risks not explained by these factors. The beta coefficients $\beta_i, \beta_j, \beta_k, \beta_l, \beta_m, \beta_s$ reflect the exposure to these respective factors, and the intercept α_i is used to obtain the GRS F-Value score.

The tables referenced from Table *3* to Table *5* present the words used to construct the dictionaries, which were subsequently utilized to generate thematic signals. These dictionaries categorize words based on their thematic relevance, enabling a structured approach to textual analysis within the context of financial markets.

Notably, the Mundane Words Dictionary contains a significantly larger set of words compared to the other dictionaries. This discrepancy arises from the inherent characteristics of the thematic grouping associated with the Mundane dictionary, which includes commonly used, neutral words that appear frequently in financial text but may not convey strong sentiment or predictive signals. The extensive word count within this dictionary reflects the broad linguistic scope of mundane terms, contrasting with the more selective nature of thematic word lists designed to capture sentiment, uncertainty, or other targeted financial attributes (Loughran and McDonald, 2011).

The use of predefined dictionaries in financial textual analysis follows established methodologies in the literature, where categorizing words into meaningful groups has been shown to improve the interpretability and predictive power of sentiment-driven investment strategies (Tetlock, 2007) (Garcia, 2013). By organizing words into structured thematic dictionaries, the methodology ensures consistency

and robustness in the generation of text-based financial signals.

Mundane Words Dictionary				
good	day	second	quarter	my
your	today	end	being	would
now	like	turn	over	again
me	during	related	business	these
only	should	future	including	then
ill	come	back	bit	some
outlook	thanks	us	were	half
high	was	todays	strong	key
want	focus	first	look	another
their	third	weve	past	im
pleased	share	progress	existing	very
early	into	where	work	support
additional	years	out	than	expect
continue	but	thing	within	need
service	across	overall	use	well
one	its	going	line	world
range	even	move	data	get
network	same	plan	able	them
actually	example	seeing	know	last
focused	next	also	about	new
part	both	two	many	much
expected	compared	team	building	give
rate	between	full	they	months
terms	every	forward	theres	let
long	process	just	opportunity	base
given	channel	help	ahead	through
take	did	change	lower	when
level	ago	grow	id	mentioned
earlier	engagement	done	success	add
excited	who	course	large	continues
so	areas	momentum	due	increased
increase	higher	looking	cash	down
period	build	things	positive	impact
go	prior	if	great	questions
question	coming	guys	start	here
because	youre	lot	couple	maybe
how	think	what	versus	youve

talk	whats	got	kind	okay
why	dont	doing	yes	talked
had	really	different	obviously	side
solution	see	pretty	big	sort
do	sure	around	few	still
thats	getting	say	feel	seen
little	something	mean	deal	guess
people	products	point	saw	using
cost	always	wanted	flow	said
probably	better	certainly	comes	applications
head	fiscal	free	expense	top
value	improvement	loss	term	already
macro	supply	group	morning	continued
verification	identity	pandemic	net	international
currency	commercial	backlog	titles	government
slide	maintenance	pipeline	security	client
businesses	software	enterprise	media	units
executed	contracts	plan	initiated	category
performance	basis	terms	spreads	systems

TABLE 3: Mundane Words Dictionary: individual words employed for the construction of the mundane themed signal

Proper Nouns Words Dictionary				
bryan	anthony	lee	jeffrey	jeff
doug	terry	tom	craig	josh
gary	ryan	daniel	dan	stephen
davis	bob	thomas	ed	scott
adam	brent	jim		

TABLE 4: Proper Noun Words Dictionary: individual words employed for the construction of the proper noun signal

Technicism Words Dictionary				
ebitda	selfserve	founder	growth	acv
oem	retention	commerce	merchant	arpu
ads	consumers	healthcare	gotomarket	holdings
operating	experience	secretary	cofounder	arr
benchmark	publisher	traffic		

TABLE 5: Technicism Words Dictionary: technicism words employed for the construction of the technicism signal

Equities scores attribution

Building on the previously established quartile-based ranking of words, we assign weights to individual words according to their relative importance, as derived from their GRS Variance. This approach emphasizes words that contribute more meaningfully to the financial signal extracted from earnings call transcripts. To ensure statistical significance, only those words associated with factors exhibiting a GRS Variance greater than 0.1 are retained. The weight assigned to each word w_i is normalized across all retained factors, as described in equation 2, ensuring that the total contribution sums to one:

$$w_i = \frac{\sigma^2_{GRS,i}}{\sum_{j \in \mathcal{F}} \sigma^2_{GRS,j}} \text{ (2)}$$

where $\sigma^2_{GRS,i}$ represents the GRS Variance of the iii-th retained factor, and F is the set of all such retained factors.

These weights are applied to compute sentiment-based and non-sentiment scores for each transcript. The main measures are defined as follows:

The Weighted Positive Score quantifies the strength of positive sentiment by summing the frequency of each positive word multiplied by its assigned weight, as defined in equation 3:

$$S^+ = \sum_{i=1}^{N^+} w_i f_i^+ . \text{ (3)}$$

The Weighted Negative Score follows the same structure, aggregating the weighted frequencies of negative words, as defined in equation 4:

$$S^- = \sum_{i=1}^{N^-} w_i f_i^- . \text{ (4)}$$

The Total Sentiment Score combines both positive and negative components to represent overall sentiment intensity, as defined in equation 5:

$$S_T^{\text{Total}} = \sum_{i \in T} w_i f_i \text{ (5)}$$

Alternatively, we define the Combined Sentiment Score to emphasize the full sentiment spectrum across both directions, as defined in equation 6:

$$S_T^{\text{PosNeg}} = S^+ + S^- \text{ (6)}$$

Finally, to capture the Normalized Sentiment Polarity, we compute the net difference between positive and negative sentiment relative to their sum, as defined in equation 7.

$$S_T^{\text{PosNegNorm}} = \frac{S^+ - S^-}{S^+ + S^-} \text{ (7)}$$

Words that do not meet the minimum GRS Variance threshold are excluded to reduce the influence of noise and low-explanatory-power terms. This GRS-based weighting scheme enhances the signal quality by increasing the statistical efficiency of the sentiment estimates.

Weighted factor construction

To construct weighted sentiment-based factors, we first categorize each of the computed sentiment scores-namely, the Weighted Positive, Weighted Negative, and Weighted Total-into quartiles based on their empirical distribution. This process facilitates systematic attribution by segmenting firms according to the relative strength of their sentiment signals. The quartile-based approach is widely used in asset pricing literature to investigate the differential impact of factor exposures across ranked groups (Fama and French, 1992) (Fama and French, 2015).

The quartiles are defined as follows: Quartile 0 captures cases with zero-weighted sentiment scores, representing transcripts devoid of sentiment-rich language. Quartiles 1 to 4 represent increasing ranges of weighted sentiment scores and are computed using empirical percentiles to ensure a balanced distribution of firms across quartiles. This methodology is consistent with portfolio sorting techniques in empirical finance, wherein securities are binned based on factor loadings or characteristics to evaluate cross-sectional return behavior (Jegadeesh and Titman, 1993) (Hou et al., 2014).

To assess the financial relevance of these sentiment-derived factors, we match each firm's monthly return data with its corresponding quartile. Grouped by quartile, we compute the average return for each month. This mirrors factor performance attribution frameworks employed in empirical asset pricing (Fama and French, 1993) (Cohen et al., 2009). The average return for firms in quartile q at time t is calculated as per equation 8:

$$R_{q,t} = \frac{1}{N_q} \sum_{i=1}^{N_q} R_{i,t} \quad (8)$$

where $R_{q,t}\#\#$ denotes the monthly average return for quartile q, $R_{i,t}$ is the return of firm iii in that quartile at time ttt, and N_q is the number of firms assigned to quartile q. This procedure enables the identification of potential return differentials driven by variations in sentiment intensity, contributing to the broader literature on sentiment-based return predictability (Tetlock, 2007) (Garcia, 2013).

Monthly average returns for each quartile are saved as part of the output to facilitate longitudinal analysis of return behavior across sentiment-defined groups. The specific quartile cutoffs used for each score are also recorded to promote methodological transparency and ensure reproducibility in follow-up robustness tests. This approach, which organizes sentiment signals into discrete groups and then evaluates return implications, is well aligned with empirical strategies in factor investing and risk-based portfolio construction (Harvey et al., 2015).

The outputs of this process, the quartile-based weighted factor scores, are then treated as independent factors. Their long-term effects on return behavior are subsequently examined in the Results section dedicated to long-term impacts.

The following subsection introduces and discusses the NLP engine selected for the implementation of our sentiment analysis methodology.

Choice of NLP engine

NLP techniques are increasingly used in financial research to extract structured insights from textual data, such as earnings call transcripts, analyst reports, and corporate filings. While domain-specific lexicons like the Loughran-McDonald (LM) word list have been widely adopted in financial studies for sentiment analysis (Loughran and McDonald, 2011), our approach employs a pre-trained transformer model, DistilBERT fine-tuned on the Stanford Sentiment Treebank (SST-2) (Sanh et al., 2019) (Socher et al., 2013). This decision is motivated by the flexibility and contextual sensitivity of transformer models, which address limitations inherent in static lexicon-based approaches.

Domain-specific lexicons, such as the LM list, are effective at capturing sentiment in structured filings like 10-K reports by classifying words into predefined positive and negative categories. However, these lexicons rely on proportional word-counting methods that assume equal weighting for all words within a category (Jegadeesh and Wu, 2013). This approach may fail to account for the context-dependent meaning of words in more nuanced, conversational settings like earnings calls. For example, terms like "debt" or "risk," classified as negative in lexicons, may convey positive sentiment in contexts where they are described as being reduced or mitigated. DistilBERT, on the other hand, dynamically interprets word meaning based on its context, making it more robust for analyzing unstructured and contextually rich

narratives (Devlin et al., 2019).

The use of pre-trained models like DistilBERT offers several advantages over lexicon-based approaches. Firstly, transformer models are trained on large, diverse corpora, enabling them to generalize across a wide variety of textual data, including financial disclosures, without requiring manual curation of domain-specific vocabularies. While previous studies have highlighted the importance of tailoring word lists to specific financial contexts (Jegadeesh and Wu, 2013) (Loughran and McDonald, 2011), such efforts are resource-intensive and risk becoming outdated as language evolves. In contrast, DistilBERT captures linguistic trends dynamically, allowing it to adapt to new terminology and expressions commonly found in earnings calls, such as those introduced by technological advancements or shifts in market conditions (Sanh et al., 2019), offering a good benchmark despite not being the latest.

Sentiment analysis in financial studies often seeks to quantify tone at the document or sentence level. The proportional weighting schemes used in lexicon-based methods assume that all terms within a category contribute equally to overall sentiment, which may not always reflect the market's interpretation (Jegadeesh and Wu, 2013). In contrast, DistilBERT assigns sentiment based on the entire context of a sentence or passage, avoiding the pitfalls of uniform term weighting. This context-sensitive approach is particularly well-suited to earnings calls, where sentiment is often conveyed through complex phrases and interactions rather than isolated words.

Although domain-specific methods provide interpretable results, they can introduce subjectivity in word classification and weight assignment. This has been mitigated by deriving term weights from market reactions (Jegadeesh and Wu, 2013), yet such methods are limited to the specific dataset and context for which they were developed. DistilBERT, however, eliminates this subjectivity by using pre-trained sentiment models that have been validated across diverse datasets (Socher et al., 2013). While the model was not originally trained on financial data, its proven performance in classifying general sentiment ensures a consistent and objective framework for evaluating earnings call transcripts.

The scalability and efficiency of transformer models also make them particularly suitable for analyzing large volumes of unstructured data. Recent work has also demonstrated the use of AI-driven NLP tools for extracting structured insights from unstructured organizational text, reinforcing the relevance of such approaches beyond finance-specific domains (Mohamed et al., 2025). Financial research often involves longitudinal studies requiring the processing of decades of textual data, as in this study. While manually updating domain-specific lexicons to account for language evolution is time-intensive, transformer models like DistilBERT require minimal adjustment once trained. Moreover, the model's ability to process text at scale ensures that results are both replicable and consistent across time periods and datasets (Devlin et al., 2019).

While domain-specific lexicons remain valuable for highly targeted analyses, the flexibility, scalability, and contextual sensitivity of transformer-based NLP models provide compelling advantages in settings that require a dynamic interpretation of text. By leveraging DistilBERT fine-tuned on SST-2, our approach combines the rigor of machine learning with the adaptability needed for analyzing complex financial narratives. This ensures robust and replicable results, while offering a methodology that is scalable for future applications.

Results And Discussion

To conduct our analysis, we construct multiple word lists, each categorized into distinct thematic sections to facilitate a structured and comparative examination of textual sentiment. Each list is developed separately, allowing for targeted analysis of specific linguistic dimensions and their respective impacts on market behaviour.

We initiate our exploration with a control list, designated as JoF_Control. This list incorporates widely accepted terms known to influence market returns, drawing primarily from the H4N-Inf word list based on the Harvard-IV-4 Psychosociological Dictionary's TagNeg file and the Fin-Neg list identified in Loughran et al. (Loughran et al., 2011).

The JoF_Control list serves as the foundation for assessing methodological refinements that are subsequently applied to other word lists. To ensure precision and reduce noise, we employ only the most relevant terms from these sources rather than utilizing the entire word corpus. This decision is guided by two considerations: first, to exclude terms that may not be inherently negative, and second, to accommodate computational constraints, which prevent us from categorizing over 2,000 words across more than 15,000 reports into positive and negative categories.

As highlighted by Loughran et al. (Loughran et al., 2011), the selected terms in our control list demonstrate a high cumulative frequency, consistently accounting for 50% or more of the total frequency within both word lists. This focus on high-frequency terms ensures the robustness and relevance of our

analysis while adhering to practical computational limits.

Next, we construct three additional word lists categorized by lexical usage rather than sentiment or inherent meaning. Each list is further refined by examining the contextual sentiment associated with individual terms. The first, the Mundane List, includes function words such as prepositions and conjunctions, which serve a structural role in language but carry little to no semantic weight. The Technicism List captures domain-specific terminology relevant to the SaaS industry, encompassing technical jargon and specialized vocabulary. Finally, the Proper Noun List comprises names of individuals, such as analysts posing questions or company executives speaking during earnings calls, reflecting the presence of identifiable entities within the transcripts.

To ensure only relevant terms are included, all terms in these lists are filtered to include only those with a GRS variance greater than 0.1, as detailed in the Methodology section. Specific entries for each list, along with their corresponding GRS variances, are provided in the annex for reference.

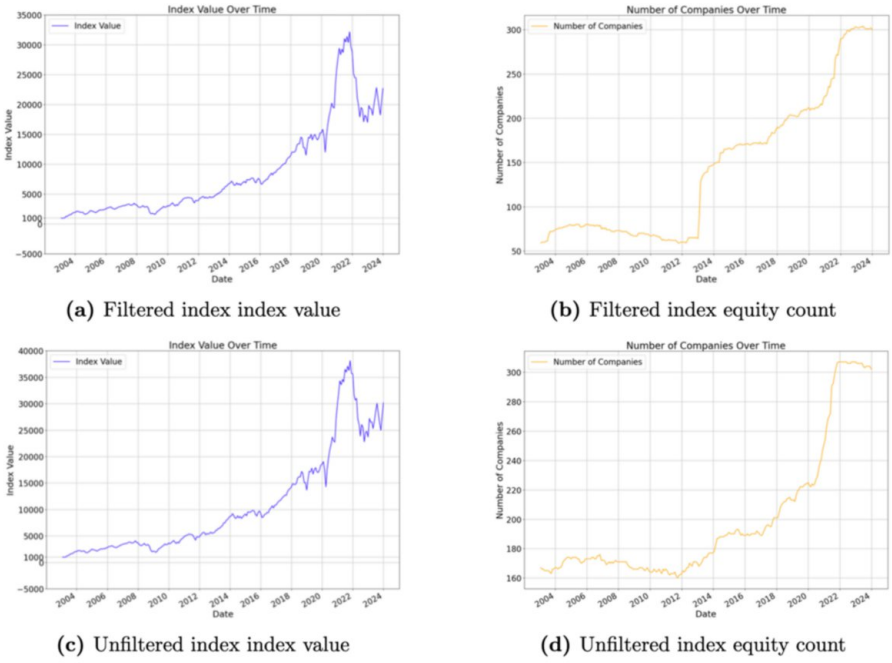
Benchmark creation

Due to the absence of a reliable SaaS-focused benchmark, the utility of existing options is limited, often introducing significant biases. Our analysis encompasses all publicly traded SaaS companies in the U.S., excluding over-the-counter (OTC) equities, during the period from January 2003 to December 2023. This contrasts sharply with the most prominent benchmark, (SaaS Capital, 2024), which includes only 97 unique companies, whereas our dataset identifies 437 (de Leon, 2025).

The limitations of SaaS Capital are further highlighted by its restrictive selection criteria, which only includes companies active as of Q3 2019 (SaaS Capital, 2024). This approach inherently excludes companies that were delisted or acquired after this date, introducing survivorship bias by capturing only those businesses that have remained successful up to that point. Additionally, the benchmark's temporal coverage from 2008 to 2024 does not align with our more extensive analysis period, significantly reducing its applicability and reliability for our comprehensive study.

The impact of survivorship bias is substantial and merits careful consideration. When survivorship bias is present, the compounded average return reaches approximately 25%, as illustrated in Panel (c) of Figure *1*. In contrast, after adjusting for this bias, the return decreases significantly to around 20%, as shown in Panel (a). This effect is further evidenced by the final index values, where Panel (a) displays a closing value of approximately 23,000 units, compared to 30,000 units in Panel (c). Figure *1* clearly visualizes this discrepancy, highlighting the considerable influence of survivorship bias on return calculations.

Moreover, the use of the SaaS Capital methodology often leads to the selection of companies only in their public stages, excluding earlier OTC stages where many companies fail. However, those that succeed from these stages often yield high returns, further emphasizing the impact of the bias. This selection process explains why more companies are counted in earlier years of our dataset, as these companies are often still in their OTC stages during our analysis period.



(a) Filtered index index value

(b) Filtered index equity count

(c) Unfiltered index index value

(d) Unfiltered index equity count

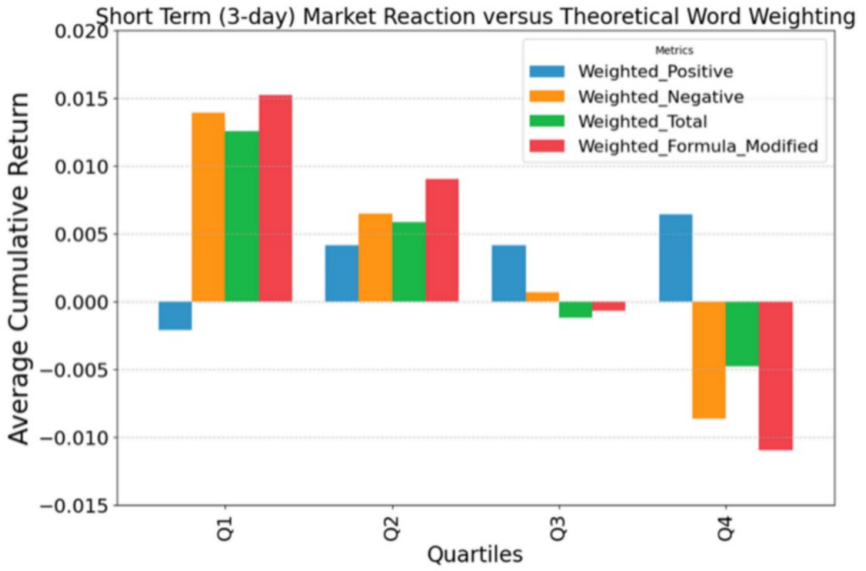
FIGURE 1: Yearly returns and number of companies over time. Panel (a) and (b) of this figure display the yearly returns and number of companies when survivorship bias is minimized by monthly verification of the companies included in the dataset. Panel (c) and (d), in contrast, shows the results without filtering

Theoretical vs experimental weightings assessment

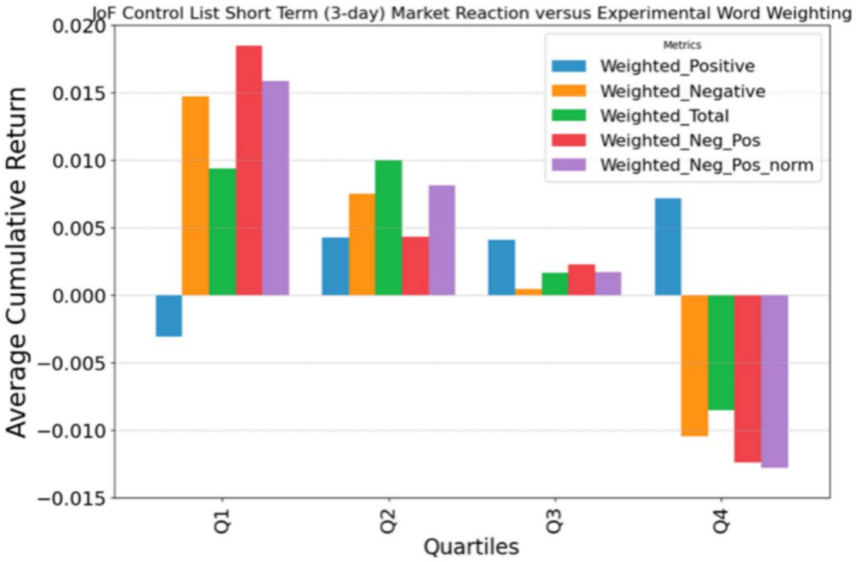
We investigate the discrepancies between theoretical (the traditional approach introduced by (Loughran and Mcdonald, 2011) and experimental weightings (the newly introduced weighthing approach based on GRS) in two distinct ways. First, we analyze the differences in returns across quartiles, which serves as an indicator of their effectiveness in classification, displayed in Figure *2*. Second, we examine the correlation between subsequent earnings calls behaviors. This approach is adopted because we are interested in a widest range possible classification that remains valid over time.

Identifying first the range behavior of different weighting schemes in Figure *2*, Panel (a), it can be observed how the consideration of sentiment is valuable. Traditionally, Weighted Total is considered only (Loughran et al., 2011), making it our benchmark for comparisons. The list against which we compare is traditionally thought of as negative words only. Despite this, we can observe how Weighted_Positive increases with increasing frequency highlighting that a liability is not always a liability, offering a new perspective to the findings of Loughran et al. (Loughran et al., 2011). This is highlighted by the fact that when removing the effect of the Weighted_Positive words from Weighted_Total yielding Weighted_Negative, the effect is enhanced, yielding a higher range than in the original benchmark. If, instead of isolating the effect, we exploit both the Positive and Negative effects, we obtain an even greater range, meaning better classification. This can be observed in Weighted_Formula_Modified, which employs a formula we develop to capture both effects.

Additionally, Panel (b) displays how the proposed new approach, where we weight word counts by their GRS variance between quartile average returns, is superior for all classifications, displaying bigger ranges. It shares the trends observed in Panel (a) but with greater ranges. Additionally, Weighted_Neg_Pos presented in the equation below, calculated per S^{PosNeg}, is the superior approach.

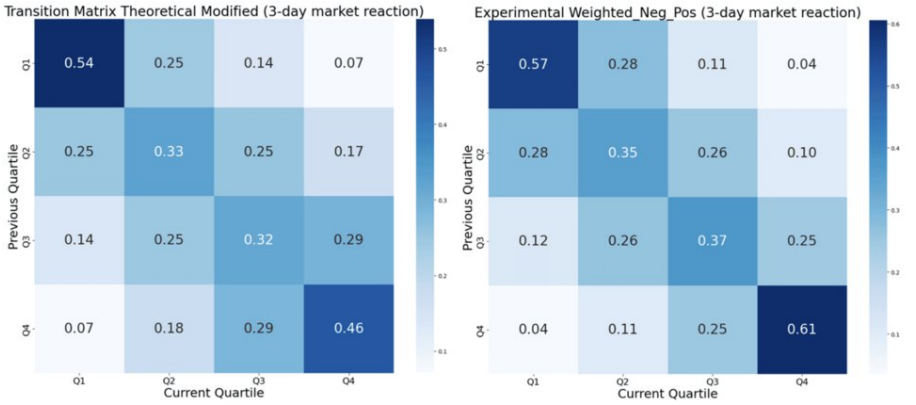


(a) Theoretical Word Weighting short term market reaction



(b) Experimental Word Weighting short term market reaction

FIGURE 2: Quartiles based on weighted word frequencies versus cumulative returns 3 days post earnings calls for the list of words JoF_Control. Panel (a) represents those weighted from a theoretical perspective, where Weighted_Formula_Modified modifies the traditional approach by constructing the weighting from the difference between the traditional weights of positive and negative words. Panel (b) employs an experimental weighting approach based on observation, as detailed in Equation (3). Weighted_Pos_Neg_norm follows Weighted_Pos_Neg but normalizes the individual scores before subtraction



(a) Transition Matrix between previous and current earning call, for the theoretical assignment approach of Weighted_Formula_Modified.

(b) Transition Matrix between previous and current earning call, for the experimental assignment approach of Weighted_Neg_Pos.

FIGURE 3: Transition matrices of cumulative returns 3 days post earnings calls, for the list of words JoF Control. Panel (a) represents those weighted from a theoretical perspective based on Weighted_Formula_Modified. Panel (b) employs an experimental weighting based on Weighted_Neg_Pos

Temporal consistency in quartile classification can be observed in the experimental weighting, which outperforms the theoretical one, displaying a higher correlation at classifying all quartiles, as per Figure *3*. Both studied effects lead to the conclusion that the experimental weighting is more effective. Consequently, it is employed in the remainder of our analysis due to its superior performance in both dimensions.

Sentiment directionality assessment

Traditionally, sentiment analysis in literature has been approached in two primary ways: by compiling a dictionary of words that are indicative of positive or negative events, often focusing predominantly on the negative, or by utilizing sophisticated NLP models to assess the overall sentiment directionality of documents or news articles. However, there has been limited research exploring whether words typically classified as negative maintain their sentiment within a positive context. In this section, we address this gap by investigating the context-dependent interpretation of positive and negative words. Additionally, we shift our focus from the conventional sentiment-oriented analysis to exploring thematic categories. Furthermore, we tailor our study to earnings calls, a domain that has been less explored compared to other types of documents, primarily due to the challenges associated with accessing the relevant transcripts.

As presented in Table *6*, the blending of positive and negative words within the Weighted_Total context notably muddles the observed effect, yielding the smallest ranges observed in all cases. This phenomenon appears to manifest in three of the four proposed word thematic groupings, highlighting that these groupings are thematic rather than sentiment-specific. In contrast, the JoF groupping, used here as a control, also exhibits a smaller range. This occurs despite its word theme being sentiment-specific, focused primarily on negative words. Consequently, it becomes evident that sentiment categorization holds value across both the control and the newly identified themes. The improvement in categorization is particularly significant in thematic groupings, enhancing precision dramatically, while the sentiment-specific groups also show a modest yet substantial enhancement.

A more detailed visualization of this effect is evident in Panel (b) of the previously presented Figure *2* and Figure *4*. These plots reveal the non-linear behavior of Weighted_Total, with a notable spike in Quartile 2, in contrast to the sentiment-categorized behavior, which typically shows a spike in Q1 followed by a continuous decline through to Q4, or vice versa. This pattern further underscores the superiority of the classification method employed.

Count	Mundane	Technicism	Proper Noun	JoF
Weighted_Positive	2.15	1.02	0.77	1.02
Weighted_Negative	1.02	0.31	0.35	2.52
Weighted_Total	0.08	0.3	0.11	1.79
Weighted_Pos_Neg	2.01	1.64	1.47	3.08
Weighted_Pos_Neg_norm	3.33	2.05	1.86	2.86

TABLE 6: Average cumulative returns ranges between difference in returns between Q1 and Q4

(a) Experimental weighting mundane list market reaction

(b) Experimental weighting technicisms list market reaction

(c) Experimental weighting proper nouns list market reaction

FIGURE 4: Quartiles based on weighted word frequencies versus cumulative returns 3 days post earnings calls for the lists of words Mundane_List, Technicism_List, and ProperNoun_List, employing the experimental approach. Panel (a) represents Mundane_List. Panel (b) displays Technicism_List. Panel (c) shows ProperNoun_List.

We now proceed to derive investment strategies based on the results obtained in previous sections.

Derived investment approach

The methodology centers around categorizing words into quartiles based on their sentiment impact, the more positive the higher the quartiles the more negative the lower. The top quartiles (Q4 and Q3) are used to establish long positions, while the lower quartiles (Q1 and Q2) are employed for short positions, as

outlined in Table 7. Specifically, if two or more quartiles fall into Q3 or Q4, a long position is taken. Conversely, if two or more quartiles fall into Q1 or Q2, a short position is adopted. The resulting investment startegy is displayed in Figure 5.

For long strategies, we recognize that acting after the event is often too late, so we utilize the outcome of the previous earnings call due to the high correlations observed in the transition matrix analysis. For short positions, we focus on the current earnings call, as previous analyses have shown that negative return days are more likely to occur after earnings calls, although the magnitude tends to be smaller.

Furthermore, as outlined in Table 8, we take long positions prior to the earnings call, based on the results of the previous call, and take short positions based on the current results, with positions being taken after the earnings call.

To reduce volatility, the portfolio allocates a maximum of 40% of the capital to any individual position. The capital is divided evenly between long and short positions when multiple signals are traded in the same day.

Item	Long strategy	Short strategy
Signals	Prev_Weighted_Pos_Neg_JoF	Weighted_Pos_Neg_JoF
	Prev_Weighted_Pos_Neg_Technicism_norm	Weighted_Pos_Neg_Technicism_norm
	Prev_Weighted_Pos_Neg_Proper_Nouns_norm	Weighted_Pos_Neg_Proper_Nouns_norm
	Prev_Weighted_Pos_Neg_Mundane_norm	Weighted_Pos_Neg_Mundane_norm
Quartiles	Q3, Q4	Q1, Q2

TABLE 7: Longed and shorted quartiles for the proposed strategy

Action	Long strategy			Short strategy		
	Premarket	During Market	Post Market	Premarket	During Market	Post Market
Entry	Price Close Day -2	Price Open Day -1	Price Open Day -1	Ask Day 0	Price Close Day 0	Ask Day 1
Exit	Price Open Day 1	Price Open Day 1	Price Open Day 1	Bid Day 2	Bid Day 2	Bid Day 2

TABLE 8: Entry and exit positions schedule, where day 0 is the day the earnings call takes place, and during market is defined as 09:30 to 15:00

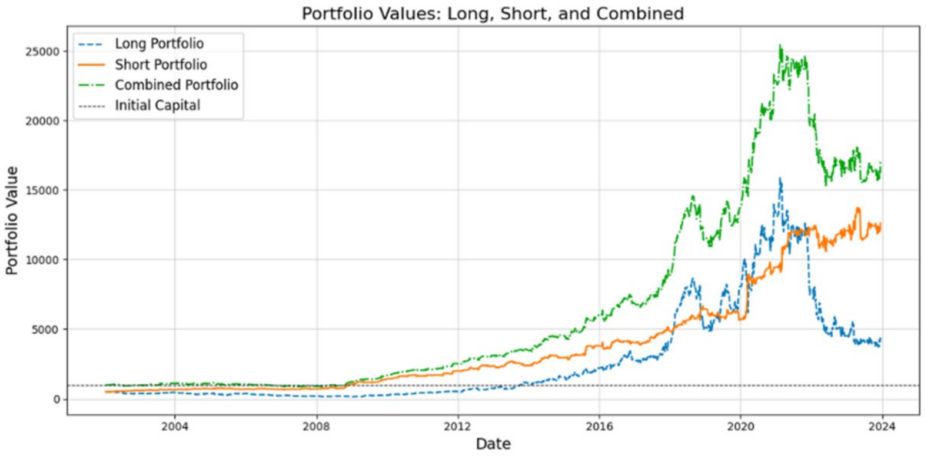


FIGURE 5: Long-, short-, and combined-signal, days-only investment backtesting, 2003 to 2024, trading only during signal days

We now present the findings from our analysis, comparing the performance of portfolios that leverage earnings call word categorization and trading signals. The portfolios span the period from 2003 to 2024, and we focus on two strategies: one that trades only on days with signals and is indexed to NASDAQ during non-signal days, and another that trades based solely on the presence of signals while staying off the market otherwise. The results highlight the efficacy of these strategies in outperforming the market while managing risk.

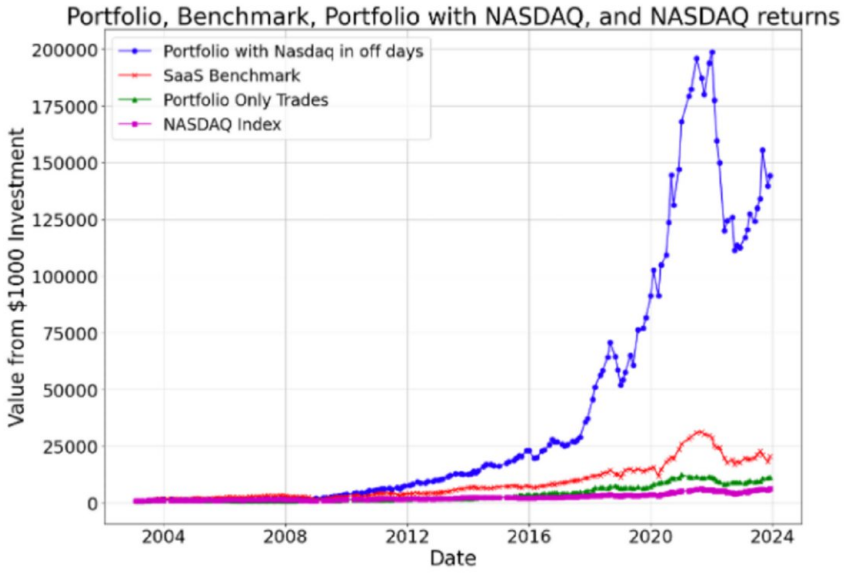


FIGURE 6: Strategy performance with and without indexing, compared against the NASDAQ and self-developed SaaS Index as benchmarks

The portfolio that is indexed to NASDAQ during off-market periods consistently outperforms both the SaaS and NASDAQ benchmark indices, as per Figure *6*. The portfolio shows a significant increase in value over time, particularly during active trading periods. This suggests that incorporating NASDAQ returns during non-active trading days enhances the overall return without exposing the portfolio to market risks during times when no signals are present. The strategy of indexing the portfolio to NASDAQ on non-signal days allows it to capture the broader market's performance, providing a stable return when no trading signals are available, and maximizing returns when signals are strong.

Metric	SaaS benchmark		NASDAQ benchmark	
	Portfolio with NASDAQ in off days	Portfolio only trades	Portfolio with NASDAQ in off days	Portfolio only trades
Benchmark average return	20.80%		15.40%	
Portfolio average return	31.30%	19%	31.30%	19%
Beta	0.73	0.14	0.73	0.19
Jensen's alpha	15.40%	13.60%	15.40%	13.70%
Sortino Ratio	1.13	1.54	1.13	1.55
Sharpe Ratio	0.89	0.72	0.9	0.72

TABLE 9: Portfolio returns metrics for our whole population; returns are rounded to the first decimal, other metrics are rounded to two decimals. Calculated with an average annual risk-free rate of 2.96% over the period

Metric	SaaS benchmark		NASDAQ benchmark	
	Portfolio with NASDAQ in off days	Portfolio only trades	Portfolio with NASDAQ in off days	Portfolio only trades
Benchmark average return	20.80%	20.80%	15.40%	15.40%
Portfolio average return	30.80%	16.18%	30.80%	16.18%
Beta	0.78	0.2	1.13	0.24
Jensen's Alpha	13.90%	9.70%	13.70%	10.30%
Sortino Ratio	2.12	1.7	2.12	1.7
Sharpe Ratio	0.87	0.64	0.87	0.65

TABLE 10: NASDAQ only equities portfolio returns metrics; returns are rounded to the first decimal, other metrics are rounded to two decimals. Calculated with an average annual risk-free rate of 2.96% over the period

Metric	SaaS benchmark		NASDAQ benchmark	
	Portfolio with NASDAQ in off days	Portfolio only trades	Portfolio with NASDAQ in off days	Portfolio only trades
Benchmark average return	21.27%		23.10%	
Portfolio average return	34.30%	11.74%	34.30%	11.74%
Beta	1.12	0.72	1.28	0.7
Jensen's alpha	10.80%	-4.40%	5.60%	-5.30%
Sortino Ratio	-	2.75	-	2.75
Sharpe Ratio	0.66	0.26	0.66	0.26

TABLE 11: Out of sample period 2019 to 2024 NASDAQ-only equities portfolio returns metrics; returns are rounded to the first decimal, other metrics are rounded to two decimals. Calculated with an average annual risk-free rate of 2.96% over the period

Performance metrics in Table *9*, Table *10* and Table *11* reveal that the portfolio with NASDAQ returns on off-market days demonstrates a notable outperformance relative to the benchmark. It achieves higher average returns, indicating the effectiveness of this strategy in capturing value during market down times while remaining unexposed to risk. The strategy also manages downside risk effectively, as evidenced by the high Sortino Ratio, which focuses on downside volatility. Although the Sharpe Ratio is below 1, which suggests that the portfolio does not fully compensate for its risk relative to traditional lower-risk investments, the superior Sortino Ratio indicates that the strategy excels in protecting against negative returns, emphasizing risk mitigation over overall volatility.

In contrast, the portfolio that only trades on days with signals demonstrates a more conservative risk profile. This strategy reduces market exposure during periods of low activity, relying on specific signals to determine trade execution. While the return is lower compared to the NASDAQ-indexed strategy, the portfolio remains highly selective in its market engagement. The lower Beta value for this portfolio indicates that it is less correlated with broader market movements, which may be advantageous during periods of market instability or when the signals are weak. The higher Sortino Ratio further highlights the portfolio's focus on managing downside risk, as it avoids market exposure during periods of low confidence.

The out-of-sample period from 2019 to 2024 reinforces the robustness of the NASDAQ-indexed strategy. We refer to this period as out of sample as it was originally not employed to identify the relevant words and their respective weights. During this period, the portfolio that incorporates NASDAQ returns on off-market days significantly outperforms the market, confirming that this strategy remains effective even in more recent, volatile market conditions. Despite a Sharpe Ratio below 1, the portfolio's ability to generate positive excess returns relative to the market risk is clear, as indicated by the strong Jensen’s Alpha values.

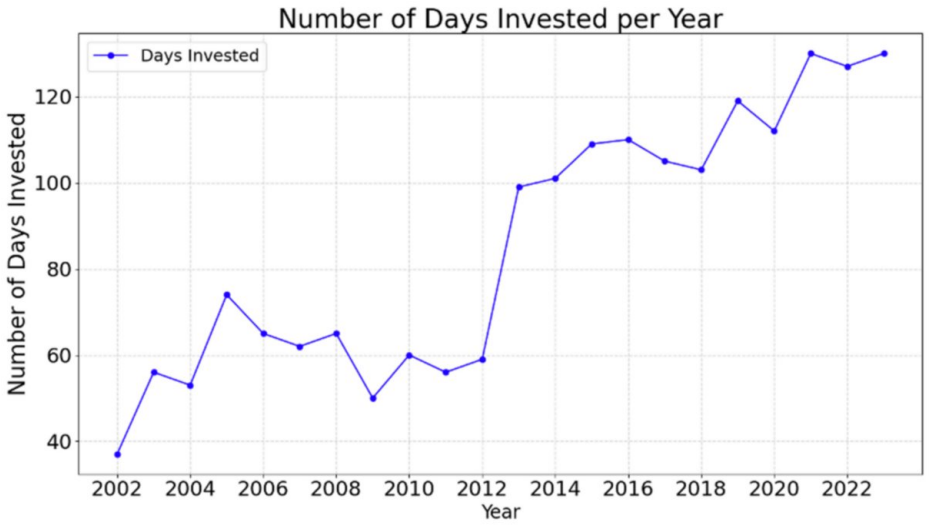


FIGURE 7: Number of days where signals occur and are acted upon (invested)

An important observation from the analysis, as shown in Figure 7, is the fluctuation in the number of days invested per year, which is influenced by the frequency of earnings calls and whether they occur on the same day or not. The number of investment days increases when a larger number of companies within the population have earnings calls during a particular period. Since the proposed strategy does not invest every day, we propose an alternative approach: indexing the portfolio during the remaining days when no trades are made. This ensures the portfolio remains active and captures market returns even on non-investment days.

Overall, the results demonstrate that incorporating NASDAQ returns during off-market days provides a strong foundation for outperforming the market. The strategy shows strong returns, and high downside protection. Although the Sharpe Ratios are suboptimal, the Sortino Ratios indicate that the portfolio excels in minimizing downside risk, making it a viable investment strategy for capturing excess returns while managing risk. The findings suggest that using the proposed signal-based trading strategy, combined with NASDAQ indexing during off-market periods, is an effective way to enhance portfolio performance, mitigate risk, and remain largely uncorrelated with the broader market.

Conclusions

This study demonstrates that the language used in software earnings calls provides predictive insights into market behavior. We introduce a context-aware linguistic model tailored to the SaaS industry, categorizing words into technicisms, proper nouns, and mundane terms. This classification captures not only sentiment but also the structural and thematic nuances that shape investor interpretation.

Our work represents the first industry-specific linguistic analysis of earnings calls. It is also the first to propose a benchmark specifically designed for publicly traded SaaS companies. This benchmark addresses survivorship bias and reflects the distinct financial and operational characteristics of the sector more accurately than general market indices. Importantly, this research spans a 21 year period, from 2003 to 2024, significantly exceeding the temporal scope of prior studies, which typically rely on shorter windows of four years or less.

The analysis, covering over four hundred companies, shows that language is an investable signal. We find that technicisms and named individuals often convey fundamental financial information, while seemingly neutral words contribute meaning through context. This insight is made possible by a novel experimental weighting methodology, based on statistical variance in returns, that captures how word frequency interacts with market dynamics.

The main contribution of this research lies in the integration of structured word-level analysis with transformer-based natural language processing. Our experimental weighting methodology significantly improves the performance of sentiment classification. The results confirm that context-sensitive linguistic features outperform traditional sentiment lexicons in identifying market-relevant language.

Compared to existing sentiment models, our approach achieves greater classification accuracy and stronger return prediction. Portfolios constructed using these linguistic signals deliver an annual compound return of approximately 30%, outperforming the benchmark average of approximately 20%.

These portfolios also maintain a lower Beta of 0.78, indicating reduced market exposure. The findings show that sentiment classification is most effective when positive and negative signals are measured together, and when context is preserved in word use.

While the study makes several contributions, it is subject to limitations. The analysis focuses on North American firms and English-language transcripts. It does not incorporate vocal tone, speaker identity, or dialogue structure, which may also affect investor responses. Future research may expand this framework to other industries, international markets, and real-time financial applications.

This work contributes to financial text analytics by combining document-level sentiment analysis with word-level structure and statistical weighting. It demonstrates that precision in language modeling can materially enhance investment strategy. As artificial intelligence continues to transform financial markets, understanding how language conveys strategic signals will be essential for those seeking to predict and outperform the market.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Jose Juan De Leon, Francesca Medda

Acquisition, analysis, or interpretation of data: Jose Juan De Leon

Drafting of the manuscript: Jose Juan De Leon

Critical review of the manuscript for important intellectual content: Jose Juan De Leon, Francesca Medda

Supervision: Francesca Medda

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The SaaS Benchmark Index is publicly available at the GitHub repository. Data on earnings call word counts used in this study are available from the corresponding author upon reasonable request.

References

1. Biddle GC, Hilary G, Verdi RS: How does financial reporting quality relate to investment efficiency?. Journal of Accounting and Economics. 2009, 48:112-31. 10.1016/j.jacceco.2009.09.001
2. Bloomfield R: Discussion of "Annual report readability, current earnings, and earnings persistence". Journal of Accounting and Economics. 2008, 45:248-52. 10.1016/j.jacceco.2008.04.002
3. Bochkay K, Hales J, Chava S: Hyperbole or reality? investor response to extreme language in earnings conference calls. The Accounting Review. 2020, 95:31-60. 10.2308/accr-52507
4. Brockman P, Li X, Price SM: Differences in conference call tones: managers vs. analysts. Financial Analysts Journal. 2015, 71:24-42. 10.2469/faj.v71.n4.1
5. Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 159:1877-901.
6. Buehlmaier MMM, Whited TM: Are financial constraints priced? Evidence from textual analysis. The Review of Financial Studies. 2018, 31:2693-728. 10.1093/rfs/hhy007
7. Bushee BJ, Gow ID, Taylor D: Linguistic complexity in firm disclosures: obfuscation or information?. Journal of Accounting Research.. 2017, 10.2139/ssrn.2375424
8. Chen JV, Nagar V, Schoenfeld J: Manager-analyst conversations in earnings conference calls. Review of Accounting Studies. 2018, 23:1315-354. 10.1007/s11142-018-9453-3
9. Cohen RB, Polk C, Vuolteenhaho T: The price is (almost) right. The Journal of Finance. 2009, 64:2739-782. 10.1111/j.1540-6261.2009.01516.x
10. Davis AK, Tama-Sweet I: Managers use of language across alternative disclosure outlets: earnings press

releases versus MD&A. SSRN. 2011, 10.2139/ssrn.1866369
11. Devlin J, Chang MW, Lee K, Toutanova K: Bert: pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics. 2019, 10.48550/arXiv.1810.04805
12. de Leon JJ: SaaS Benchmark Index [Data set]. GitHub, 2025.
13. Fama EF, French KR: A five-factor asset pricing model. Journal of Financial Economics. 2015, 116:1-22. 10.1016/j.jfineco.2014.10.010
14. Fama EF, French KR: Common risk factors in the returns on stocks and bonds. Journal of Financial Economics. 1993, 33:3-56. 10.1016/0304-405x(93)90023-5
15. Fama EF, French KR: The cross-section of expected stock returns. The Journal of Finance. 1992, 47:427-65. 10.1111/j.1540-6261.1992.tb04398.x
16. Feldman R, Govindaraj S, Livnat J, Segal B: Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies. 2009, 15:915-53. 10.1007/s11142-009-9111-x
17. Feng L: The The information content of forward-looking statements in corporate filings a naïve Bayesian machine learning approach. Journal of Accounting Research. 2010, 48:1049-102. 10.1111/j.1475-679x.2010.00382.x
18. Fu X, Wu X, Zhang Z: The information role of earnings conference call tone: evidence from stock price crash risk. Journal of Business Ethics. 2021, 173:643-60. 10.1007/s10551-019-04326-1
19. Garcia D: Sentiment during Recessions. The Journal of Finance. 2013, 68:1267-300. 10.1111/jofi.12027
20. Garcia D, Hu X, Rohrer M: The colour of finance words. Journal of Financial Economics. 2023, 147:525-49. 10.1016/j.jfineco.2022.11.006
21. Gibbons MR, Ross SA, Shanken J: A test of the efficiency of a given portfolio. Econometrica. 1989, 57:1121-152. 10.2307/1913625
22. Haddi E, Liu X, Shi Y: The role of text pre-processing in sentiment analysis. Procedia Computer Science. 2013, 17:26-32. 10.1016/j.procs.2013.05.005
23. Harvey CR, Liu Y, Zhu H: ... And the cross-section of expected returns. The Review of Financial Studies. 2015, 29:5-68. 10.1093/rfs/hhv059
24. Hou K, Xue C, Zhang L: Digesting anomalies: an investment approach. The Review of Financial Studies. 2014, 28:650-705. 10.1093/rfs/hhu068
25. Henry E: Market reaction to verbal components of earnings press releases: event study using a predictive algorithm. Journal of Emerging Technologies in Accounting. 2006, 3:1-19. 10.2308/jeta.2006.3.1.1
26. Huang X, Teoh SH, Zhang Y: Tone management. The Accounting Review. 2014, 89:1083-113. 10.2308/accr-50684
27. Jegadeesh N, Titman S: Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance. 1993, 48:65-91. 10.1111/j.1540-6261.1993.tb04702.x
28. Jegadeesh N, Wu D: Word power: A new approach for content analysis. Journal of Financial Economics. 2013, 110:712-29. 10.1016/j.jfineco.2013.08.018
29. Kearney C, Liu S: Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis. 2014, 33:171-85. 10.1016/j.irfa.2014.02.006
30. Loughran T, McDonald B: Textual analysis in accounting and finance: a survey. Journal of Accounting Research. 2016, 54:1187-230. 10.1111/1475-679x.12123
31. Loughran T, McDonald B: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance. 2011, 66:35-65. 10.1111/j.1540-6261.2010.01625.x
32. Loughran T, McDonald B: Measuring readability in financial disclosures. The Journal of Finance. 2014, 69:1643-671. 10.1111/jofi.12162
33. Mayew WJ, Venkatachalam M: The power of voice: managerial affective states and future firm performance. The Journal of Finance. 2012, 67:1-43. 10.1111/j.1540-6261.2011.01705.x
34. Mohamed HM, Matimbwa H, Banzi J: Unveiling the potential of artificial intelligence in human resource management: A systematic review of adoption strategies, challenges and future directions. Cureus Journal of Business and Economics. 2025, 10.7759/s44404-025-03698-0
35. Price SM, Doran JS, Peterson DR, Bliss BA: Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance. 2012, 36:992-1011. 10.1016/j.jbankfin.2011.10.013
36. Refinitiv Eikon: Event Screener - Earnings call transcripts and financial screening data [Proprietary database]. Refinitiv. 2024,
37. SaaS Capital. (2024). Accessed: September 30, 2025: https://www.saas-capital.com/the-saas-capital-index/#:~:text=Public%20company%20data%20is%20the,play%2C%20B2B%2C%20Sa....
38. Sanh V, Debut L, Chaumond J, Wolf T: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.. ArXiv. 2019, 10.48550/arXiv.1910.01108
39. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C: Recursive deep models for semantic compositionality over a sentiment treebank. Conference on Empirical Methods in Natural Language Processing. 2013, 1631-642.
40. Sun F, Belatreche A, Coleman S, McGinnity TM, Li Y: Pre-processing online financial text for sentiment classification: a natural language processing approach. IEEE Computational Intelligence for Financial Engineering and Economics. 2014, 10.13140/2.1.3554.9443
41. Sun C, Qiu X, Xu Y, Huang X: How to fine-tune bert for text classification?. Chinese computational linguistics: 18th China national conference, CCL 2019. Kunming, China; 2019. 18-20,. 10.48550/arXiv.1905.05583
42. Tetlock PC, Saar-Tsechansky M, Macskassy S: More than words: quantifying language to measure firms fundamentals. The Journal of Finance. 2008, 63:1437-467. 10.1111/j.1540-6261.2008.01362.x
43. Todd A, Bowden J, Moshfeghi Y: Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. Intelligent Systems in Accounting, Finance and Management. 2024, 31:e1549. 10.1002/isaf.1549
44. Tetlock PC: Giving content to investor sentiment: the role of media in the stock market. The Journal of Finance. 2007, 62:1139-168. 10.1111/j.1540-6261.2007.01232.x
45. Wharton Research Data Services (WRDS): CRSP Stock/Security Files - Market returns data [Proprietary

database]. University of Pennsylvania, 2024.