Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Integrating a Pipeline of Diverse Data Sources to estimate CEO Overconfidence

Autoría: Juárez, Lucio Ignacio
Año: 2025

UNIVERSIDAD
TORCUATO DI TELLA

Master's in Management + Analytics

Integrating a Pipeline of Diverse Data Sources to estimate CEO Overconfidence

Author: Lucio Ignacio Juárez

Advisors: Fabiana Penas and Juan José Miranda Bront

Abstract

This thesis presents the design and implementation of a data-intensive framework to estimate CEO overconfidence by integrating multiple data sources and analytical methods. Drawing from traditional sentiment-based approaches and recent advances in generative artificial intelligence (GenAI), the project develops a dynamic and scalable index to classify CEO behavior based on press articles, structured financial disclosures, and contextual analysis. The research contributes both conceptually and practically by replicating existing overconfidence indicators—such as the Conf(Press) index—and extending them through natural language processing techniques that account for nuance, context, and industry-specific factors.

The methodology combines structured data acquisition from sources like ProQuest's TDM Studio, EBSCO, and The New York Times with sentiment scoring (VADER), keyword-based classifiers, and a GenAI-powered prompt framework. By applying these techniques to over 7,000 curated CEO-related articles, the thesis constructs a CEO Overconfidence Index that enables comparative analysis across sectors, particularly between innovation-driven and traditional industries. The resulting data product captures how overconfidence varies over time and in response to events, revealing both the limitations of static keyword methods and the added value of contextual AI models.

Ultimately, this work contributes to the field of behavioral corporate finance by offering a novel pipeline to estimate executive psychological traits from textual data. It also provides a governance-relevant tool for investors, analysts, and policymakers to identify behavioral risk factors in leadership. While GenAI adds adaptability and interpretive depth, the thesis emphasizes that its primary value lies in the integration of classic and emerging methods into a unified, sector-aware overconfidence framework.

Resumen

Esta tesis presenta el diseño e implementación de un marco analítico intensivo en datos para estimar el exceso de confianza de los CEOs, integrando múltiples fuentes de información y métodos de análisis. Combinando enfoques tradicionales basados en sentimiento con técnicas recientes de inteligencia artificial generativa (GenAI), el trabajo desarrolla un índice dinámico y escalable que clasifica el comportamiento de los CEOs a partir de artículos de prensa, reportes financieros estructurados y análisis contextual. El proyecto aporta tanto en términos conceptuales como prácticos, al replicar indicadores existentes de overconfidence —como el índice Conf(Press)— y extenderlos mediante técnicas de procesamiento de lenguaje natural que incorporan matices, contexto y diferencias sectoriales.

La metodología combina la recopilación estructurada de datos provenientes de fuentes como TDM Studio de ProQuest, EBSCO y The New York Times, con métodos de análisis de sentimiento (VADER), clasificadores basados en palabras clave y un sistema de prompts desarrollado con GenAI. Estos métodos se aplicaron a más de 7.000 artículos curados relacionados con CEOs, permitiendo construir un índice de exceso de confianza ejecutiva que habilita comparaciones entre industrias, especialmente entre sectores innovadores y tradicionales. El producto final permite observar cómo este sesgo varía en el tiempo y frente a eventos relevantes, evidenciando tanto las limitaciones de los métodos clásicos como el valor agregado de modelos basados en contexto.

En última instancia, esta tesis contribuye al campo de las finanzas conductuales corporativas al ofrecer un pipeline novedoso para estimar rasgos psicológicos ejecutivos a partir de datos textuales. Asimismo, proporciona una herramienta útil para inversores, analistas y responsables de gobierno corporativo, al permitir la identificación de factores de riesgo conductuales en el liderazgo. Si bien GenAI aporta adaptabilidad y profundidad interpretativa, el valor central de este trabajo reside en la integración coherente de métodos clásicos y emergentes dentro de un marco unificado y sensible al contexto sectorial.

Table of Contents

1. Introduction

Chief Executive Officers (CEOs) play a central role in corporate strategy, shaping firm direction through high-stakes decisions in innovation, risk-taking, and acquisitions. However, the same confidence that fuels visionary leadership can also result in overconfidence—an inflated belief in one's ability to forecast outcomes, manage uncertainty, and outperform peers. CEO overconfidence has been associated with both breakthrough innovations and costly strategic failures, making it a key variable in understanding firm behavior and performance (Malmendier and Tate, 2005; Hirshleifer, Low, and Teoh, 2012).

Overconfidence is not merely about taking bold risks. It is a cognitive bias that affects how executives perceive their abilities, the probability of success, and the controllability of uncertain environments. According to Moore and Healy (2008), overconfidence can manifest in three distinct ways:

- Overestimation: believing one's abilities or strategies are better than they are.
- Overplacement: believing one is more skilled than peers.
- Overprecision: holding excessive certainty in forecasts and predictions.

While overconfidence can drive aggressive expansion, ambitious mergers, and technological breakthroughs, it also increases the probability of corporate miscalculations, financial mismanagement, and even governance failures. For instance, visionary CEOs like Steve Jobs and Elon Musk have demonstrated the power of confidence in leading companies to extraordinary innovation. Yet, historical examples like the collapse of Enron under Kenneth Lay, the downfall of Lehman Brothers under Richard Fuld, and BlackBerry's failure to adapt under its co-CEOs illustrate the risks when overconfidence is unchecked.

Given its significant impact, how can CEO overconfidence be measured accurately across industries? Several methodologies have been proposed in the literature:

- Stock-Option-Based Measures: Malmendier and Tate (2008) define overconfident CEOs as those who delay exercising in-the-money stock options, believing their firm's stock will continue to rise. This behavioral pattern reflects an overestimation of future returns and reluctance to diversify personal wealth.
- Earnings Forecast-Based Measures: Overconfident CEOs often overestimate future firm performance, leading to optimistic earnings forecasts that later prove inaccurate (Zhao and Ziebart, 2017). This method serves as a quantitative proxy for measuring the impact of overconfidence on financial decision-making.
- Press-Based Sentiment Analysis: Hribar and Yang (2016) develop "Conf(Press)," a sentiment-based measure of CEO overconfidence that analyzes media coverage and descriptors of executive behavior. A CEO frequently portrayed as confident and optimistic—compared to cautious or risk-averse—receives a higher overconfidence score.
- AI-Driven Approaches: Advances in Natural Language Processing (NLP) and Machine Learning (ML) enable sentiment and speech pattern analysis for CEOs, offering an alternative to static sentiment measures (Kaplan, Sørensen, and Zakolyukina, 2021).
- Emerging AI Techniques: Recent advancements in Generative AI (GenAI) models have introduced context-aware sentiment analysis, allowing for a more nuanced evaluation of CEO overconfidence in real-time speech and press coverage—marking a promising direction for executive behavior analysis.

While existing methods provide valuable insights, they remain limited in their ability to capture the nuanced and evolving nature of CEO behavior. Stock-option and earnings-based metrics lack real-time applicability, while press-based approaches depend on predefined sentiment categories that may not fully capture contextual overconfidence.

This thesis develops and evaluates a data-intensive application for measuring CEO overconfidence by integrating multiple data sources using an AI-driven analytical pipeline. The research aims to:

1. Design a CEO Overconfidence Index that combines press-based sentiment analysis and AI-driven textual analysis.
2. Compare traditional and AI-enhanced methods, assessing whether a Generative AI model provides a more accurate and scalable measurement framework.
3. Examine sectoral differences in CEO overconfidence between innovation-driven and traditional industries.

The data pipeline constructed in this research aggregates and processes financial news, corporate reports, and executive speeches, applying natural language processing (NLP), generative artificial intelligence (GenAI), and sentiment analysis to quantify CEO overconfidence dynamically. This approach builds upon prior methodologies while addressing their limitations in scale, adaptability, and contextual accuracy.

By integrating diverse data sources into a single analytical framework, this study contributes to:

- Corporate Governance Research: Offering a real-time, data-driven tool for investors, boards, and analysts to monitor executive decision-making biases.
- Behavioral Corporate Finance: Extending Malmendier and Tate's (2005, 2008) overconfidence frameworks with context-aware methods.
- Sectoral Insights: Unveiling differences in CEO overconfidence across industries, helping firms design better risk mitigation strategies.

The following sections of this thesis are structured as follows:

- Chapter 2 provides a literature review on CEO overconfidence, its effects, and existing measurement techniques.
- Chapter 3 outlines the data sources, AI methodology, and NLP techniques used to construct the CEO Overconfidence Index.
- Chapter 4 presents the findings, including a comparative analysis of AI-based vs. traditional overconfidence measures.
- Chapter 5 discusses implications for corporate governance, investment strategies, and potential future applications.
- Chapter 6 concludes with key takeaways and areas for future research.

By refining the measurement of CEO overconfidence through a data-driven, AI-enhanced approach, this research aims to bridge behavioral finance theory and real-world executive decision-making, providing valuable insights for both academia and industry.

2. Literature Review

2.1. How the Literature Measures CEO Overconfidence

A wide range of methods have been developed to measure CEO overconfidence, falling broadly into two categories: quantitative models based on observable financial behavior, and qualitative approaches that analyze language, perception, or sentiment.

Among the quantitative measures, stock-option-based models are one of the most established. Malmendier and Tate (2008) propose that CEOs who delay exercising in-the-money stock options, assuming continued stock appreciation, can be classified as overconfident. Another widely used method is earnings forecast optimism, where Zhao and Ziebart (2017) find that overconfident CEOs frequently issue optimistic earnings forecasts that consistently miss expectations. More recently, AI-Based Analysis has introduced Bayesian models to estimate CEO overconfidence probabilities, integrating textual analysis and firm characteristics (Hatoum et al., 2022).

On the qualitative side, press-based sentiment analysis has emerged as a valuable proxy. Hribar and Yang (2016) introduce the Conf(Press) index, which classifies CEOs as overconfident based on the frequency of media descriptors such as "confident," "bold," or "optimistic." Complementing this, Kaplan, Sørensen, and Zakolyukina (2021) leverage structured behavioral interviews to assess the gap between self-perception and external evaluation, offering an alternative lens on executive self-assessment biases.

While each method captures important dimensions of CEO overconfidence, no single approach is sufficient on its own. Consequently, recent research trends emphasize the integration of multiple data sources and methods—financial, linguistic, and contextual—to improve robustness and interpretability in overconfidence measurement.

2.2. The Dual Nature of CEO Overconfidence

CEO overconfidence has been widely studied in behavioral finance and corporate governance, with researchers highlighting both its constructive and destructive effects on firm behavior. On the one hand, overconfidence can act as a strategic asset that fuels innovation and leadership assertiveness. On the other, it can distort risk assessment, impair capital allocation, and weaken governance mechanisms. This section explores both dimensions of CEO overconfidence and the evolution of methods used to measure it.

In certain industries, particularly those characterized by uncertainty and fast-paced innovation, CEO overconfidence is often interpreted as a virtue. Overconfident executives are more likely to act decisively, pursue ambitious projects, and challenge conventional wisdom—traits that can differentiate their firms competitively. Studies suggest that such behaviors can create long-term value under the right conditions. However, when confidence turns into systematic bias—especially in capital-intensive, low-innovation contexts—the same traits can lead to value destruction. Thus, overconfidence must be analyzed as a double-edged phenomenon whose impact is shaped by organizational, financial, and industrial context.

2.3. Positive Effects

CEO overconfidence is widely cited as a driver of innovation and bold strategic decisions, with overconfident executives more willing to pursue ambitious projects that risk-averse leaders might reject.

Hirshleifer, Low, and Teoh (2012) find that firms led by overconfident CEOs invest more in high-risk innovation, leading to more patent citations and greater innovative success. Similarly, Galasso and Simcoe (2011) show that overconfident CEOs push their firms into new technological directions, increasing citation-weighted patents, particularly in competitive industries. Additionally, Goel and Thakor (2008) argue that overconfidence helps firms commit to investment decisions that risk-averse executives might avoid, reducing decision-making inertia.

Beyond innovation, overconfidence is linked to stronger leadership influence and market positioning. Phua, Tham, and Wei (2018) provide evidence that overconfident CEOs enhance stakeholder commitment, as employees and suppliers perceive their strong belief in the firm's prospects as a sign of stability and long-term reliability. Gervais, Heaton, and Odean (2011) suggest that overconfident CEOs gain stronger support from shareholders and boards, enabling them to pursue more ambitious corporate strategies. Furthermore, Petit and Bollaert (2012) discuss how charismatic, overconfident CEOs influence media portrayals, reinforcing investor sentiment and shaping public perceptions of their firm's competitive strength.

2.4. Negative Effects

Despite its potential benefits, CEO overconfidence frequently leads to excessive risk-taking, financial mismanagement, and governance issues, often with long-term detrimental effects on firm performance. In Mergers and Acquisitions (M&A) transactions, overconfident executives often overestimate their ability to extract synergies, leading to flawed decision-making.

Malmendier and Tate (2008) find that overconfident CEOs are significantly more likely to overpay for acquisitions, particularly when using internal financing. The market reaction to such deals is more negative than for rational CEOs. Their findings align with Roll's (1986) Hubris Hypothesis, which suggests that CEOs suffering from hubris believe they are uniquely capable of making an acquisition successful, leading them to bid too high. Zhao and Ziebart (2017) find that bond yield spreads are significantly higher for firms led by overconfident CEOs. This reflects market concerns about excessive optimism and perceived default risk.

In addition to M&A missteps, overconfidence also affects financial management and corporate governance. Malmendier and Tate (2005, 2008) show that investment by firms with overconfident CEOs is more sensitive to cash flow, as they avoid external financing, even when it would be optimal. Hribar and Yang (2016) introduce the Conf(Press) Index, a media-based measure of overconfidence, showing a significant correlation between media portrayals of confidence and excessive capital expenditures by CEOs. Lastly, Petit and Bollaert (2012) warn that weakened board oversight allows overconfident CEOs to push forward high-risk strategies with reduced scrutiny, increasing governance conflicts.

Given these risks, accurately measuring CEO overconfidence is essential to understanding its influence on corporate behavior. Researchers have developed both quantitative and qualitative methods to assess and quantify this trait, offering complementary insights into CEO behavior.

2.5. Sectoral Nuance: Innovative vs. Traditional Industries

The impact of CEO overconfidence varies significantly based on industry type, influencing how firms benefit—or suffer—from executive biases. In highly innovative industries, overconfidence can be an asset, driving technological advancements, whereas in more traditional, capital-intensive firms, it often results in financial overextension and increased risk exposure.

In innovation-intensive firms, overconfidence can be beneficial. Hirshleifer, Low, and Teoh (2012) show that overconfidence enhances R&D productivity, leading to higher patent output and breakthrough innovations. Likewise, Galasso and Simcoe (2011) demonstrate that overconfident CEOs in highly competitive markets leverage their risk appetite to drive technological advancements, improving firm valuation.

However, in traditional, capital-intensive industries, the risks of CEO overconfidence become more pronounced. Zhao and Ziebart (2017) find that in industries reliant on stable cash flows, CEO overconfidence raises borrowing costs, making capital allocation riskier. Malmendier and Tate (2008) conclude that in low-innovation sectors, overconfident CEOs are more likely to engage in destructive M&A strategies, depleting shareholder value.

By synthesizing these insights, this study develops a comprehensive CEO Overconfidence Index, leveraging AI-driven sentiment analysis to refine existing methodologies and offer a more accurate, context-sensitive assessment of CEO decision-making behaviors.

3. Theoretical Framework

Understanding CEO overconfidence and its implications for corporate decision-making requires an interdisciplinary theoretical foundation that integrates insights from behavioral finance, organizational psychology, strategic management, and data science. Overconfidence, as a cognitive bias, reflects an exaggerated belief in one's own capabilities, the accuracy of one's judgments, and the likelihood of favorable outcomes, even in situations marked by significant uncertainty (Malmendier and Tate, 2008; Moore and Healy, 2008).

Following Moore and Healy (2008), overconfidence can be decomposed into three interrelated dimensions:

- Overestimation, or the tendency to believe one's performance or skill is greater than it truly is, often manifesting in unrealistic strategic ambition.
- Overplacement, whereby individuals consider themselves more competent than their peers, potentially underestimating external threats or rivals.
- Overprecision, reflected in excessive confidence in the accuracy of one's information, leading to underappreciation of uncertainty and risk.

These dimensions have been shown to shape CEO behavior in high-stakes decisions. Malmendier and Tate (2005, 2008, 2011) demonstrate that overconfident CEOs often delay exercising in-the-money stock options, avoid external financing—even when it would be optimal—and engage in overly aggressive investment strategies. These behaviors suggest a consistent psychological pattern that affects firm capital structure and investment dynamics.

In the context of M&A, CEO overconfidence becomes particularly pronounced. Roll's (1986) Hubris Hypothesis points out that managers often overestimate the synergies their firm can generate through acquisitions, leading to overbidding and eventual value destruction. Malmendier and Tate (2008) reinforce this view, finding that overconfident CEOs are significantly more likely to engage in acquisitions—especially when self-financed—and that the market reacts more negatively to their deals. Extending this analysis, Zhao and Ziebart (2017) show that overconfident CEOs tend to increase a firm's cost of debt by projecting persistent optimism in earnings forecasts, which raises perceived credit risk among bondholders. This phenomenon is especially detrimental in capital-intensive, low-innovation sectors, where cost of capital directly constraints operational flexibility.

At the organizational level, Kaplan, Sørensen, and Zakolyukina (2021) add another layer of nuance. Their behavioral assessment framework shows that overconfident CEOs often discount external feedback and over-rely on internal conviction, reinforcing strategic misalignment and complicating governance oversight. These tendencies weaken the corrective mechanisms that boards and stakeholders use to realign strategy with market conditions.

From an external perception standpoint, Hribar and Yang (2016) introduce the Conf(Press) index, a press-based proxy for overconfidence that measures the frequency of positive descriptors—such as "confident," "bold," or "visionary"—in media coverage of CEOs. Their findings demonstrate a robust link between media sentiment and excessive capital expenditures, suggesting that how CEOs are portrayed can influence, and be influenced by, overconfident behavior. This introduces a feedback loop between executive identity, media framing, and firm-level decisions.

Taken together, these studies illustrate the complex psychological and organizational dynamics of CEO overconfidence. However, they also reveal a broader methodological challenge: How can overconfidence be reliably measured across firms and industries?

This challenge has motivated researchers to pursue both quantitative and qualitative approaches. Stock-option behavior and earnings forecast optimism serve as behavioral proxies, but they depend on data availability and lag real-time decisions. Sentiment analysis, such as Conf(Press), offers more scalable monitoring but is sensitive to context, wording, and narrative framing. Recognizing the limitations of single-method approaches, scholars increasingly advocate for integrative frameworks that combine financial indicators, language-based measures, and organizational signals.

In this context, data-intensive methodologies emerge as promising tools for capturing the multifaceted nature of CEO overconfidence. Davenport and Kudyba (2016) define a data product as a structured combination of analytics and data pipelines designed to produce actionable insights. In this thesis, the CEO Overconfidence Index is conceived as such a data product: combining textual data from press articles, computational methods (sentiment analysis, classification), and business rules (industry segmentation) to yield interpretable, behavioral indicators.

Dehghani (2022) further emphasizes that a robust data product must be standardized, repeatable, and context-sensitive. These qualities are essential when estimating a latent variable like overconfidence, which cannot be captured by any single indicator or event. The index proposed in this thesis is designed with these principles in mind, enabling scalability across firms and adaptability to diverse contexts.

To achieve this, the framework draws on the foundational principles of data-intensive applications, as outlined by Kleppmann (2017). These include large-scale text processing, semantic indexing, and batch analytics—all of which are necessary for handling unstructured CEO-related content at scale. The core methodological tools include natural language processing (NLP) and sentiment analysis, which enable the extraction of meaning from qualitative data sources such as media coverage and executive communications. These tools allow researchers and practitioners to decode subtle linguistic cues that traditional financial ratios or disclosures cannot capture.

Additionally, this thesis incorporates Generative Artificial Intelligence (GenAI) as a third layer of analysis. GenAI models—capable of generating context-aware responses and interpreting nuanced sentiment—can process and classify CEO-related narratives with greater semantic accuracy than rule-based lexicons alone. For example, rather than merely flagging the word "confident," a GenAI model can evaluate whether a CEO's tone reflects justified conviction or inflated self-regard, based on broader contextual cues.

This incorporation of GenAI is not intended to replace traditional models but to complement them by providing an adaptive, scalable layer of interpretation. When paired with established sentiment metrics like Conf(Press) and earnings forecast behavior, GenAI-based classification enriches the model's sensitivity to tone, context, and industry nuance.

In sum, this theoretical framework brings together behavioral theory, organizational analysis, and advanced analytics to support a novel, multidimensional approach to measuring CEO overconfidence. By grounding the CEO Overconfidence Index in both psychological theory and computational methodology, the framework ensures that the measure is both empirically robust and practically usable. It provides a scalable structure through which boards, investors, and researchers can assess and monitor executive overconfidence.

Thesis Statement and Hypotheses

This thesis proposes that CEO overconfidence can be systematically estimated from press-based textual data through a hybrid approach that combines traditional sentiment analysis methods—such as Conf(Press) and VADER—with Generative Artificial Intelligence (GenAI) models. This multi-layered framework enables the construction of a dynamic, context-sensitive, and interpretable CEO Overconfidence Index that adapts to sectoral differences and enhances the detection and monitoring of behavioral bias across industries. The core claim is that overconfident behavior is not uniformly expressed across firms but varies by context, and that this variation can be captured and analyzed at scale using data-intensive techniques.

Based on the theoretical framework developed in this chapter, and in anticipation of the research questions presented in Chapter 4, the following hypotheses are proposed:

- **H1:** GenAI-based sentiment models capture CEO overconfidence in press articles with greater contextual sensitivity and accuracy than traditional keyword-based or lexicon-polarity methods such as Conf(Press) and VADER.

- **H2:** CEO overconfidence manifests with greater frequency and intensity in innovation-driven sectors (e.g., high R&D, rapid technological change) than in traditional or capital-intensive industries.

- **H3:** Composite overconfidence scores that integrate GenAI outputs and traditional sentiment measures are positively correlated with firm behaviors associated with strategic risk and innovation, such as increased R&D investment and aggressive capital allocation.

4. Research Questions

This thesis seeks to construct a comprehensive, data-driven CEO Overconfidence Index by integrating traditional press-based sentiment analysis with emerging Generative AI (GenAI) classification techniques. In doing so, it explores how CEO overconfidence varies across industry contexts—particularly between innovation-intensive and traditional sectors—and how different measurement approaches relate to observable firm outcomes.

Specifically, the study addresses the following questions:

1. How do traditional press-based sentiment analysis methods and Generative AI (GenAI) techniques differ in their ability to capture CEO overconfidence? What are their respective strengths and limitations?

   This question examines methodological contrasts between established sentiment analysis approaches (e.g., Conf(Press)) and GenAI-based models, focusing on their contextual sensitivity, scalability, and interpretability.

2. In what ways does CEO overconfidence manifest differently in innovation-driven industries compared to traditional, capital-intensive sectors?

   This question investigates whether the expression and impact of CEO overconfidence vary across sectors, as captured by the Overconfidence Index. It considers behavioral indicators such as media sentiment, strategic posture, and investment tendencies.

3. How do the results from traditional sentiment-based methods compare to those derived from GenAI in explaining firm-level patterns, such as investment intensity and innovation orientation?

   This question evaluates the extent to which each overconfidence measurement approach aligns with observable organizational behaviors, particularly capital allocation and innovation activity. It contributes to assessing the practical utility of multi-method sentiment models in applied corporate analysis.

By addressing these questions, the thesis contributes a multi-method framework for estimating CEO overconfidence and assessing its implications across sectors. The results aim to support a more nuanced understanding of executive decision-making and offer tools for investors, analysts, and governance professionals seeking to monitor behavioral risk in leadership.

5. Methodology

5.1. Research Design

This thesis employs an exploratory research design to investigate CEO overconfidence as a latent behavioral trait expressed through textual signals. The primary goal is to develop a CEO Overconfidence Index using a multi-method sentiment analysis pipeline, capable of classifying press articles into confidence-based categories and interpreting overconfidence across sectors. This design is well-suited to addressing questions that involve uncovering patterns in qualitative data without manipulating variables or establishing causality. The research places particular emphasis on sectoral differences, drawing comparisons between innovative and traditional industries, and explores how textual signals align with firm characteristics and behavioral patterns.

Exploratory research is appropriate here because overconfidence—particularly in CEOs—is difficult to measure directly. Instead, this approach allows for triangulation between methods (keyword sentiment, polarity, AI interpretation) and dimensions (firm, sector, time). It also enables theory-building by connecting academic frameworks with practical applications for investors, analysts, and corporate stakeholders.

5.2. Data Collection Strategy

Data were collected from three main sources: (1) ProQuest TDM Studio, (2) the EBSCO Information Services API, and (3) the New York Times public archive using web scraping. These sources were selected for their combination of reputational reliability, business relevance, and breadth of coverage. The collected corpus includes over 11,000 articles spanning from 2000 to 2023, all of which explicitly reference CEOs or firm-level strategic decisions. After initial scraping, each article underwent language standardization and duplicate removal.

ProQuest TDM Studio enabled the collection of a wide array of historical financial journalism, including specialized content from outlets like *The Economist*. EBSCO's API was used to supplement this corpus with targeted queries focused on CEO profiles, interviews, and business commentary. Articles were also retrieved from the New York Times using a structured scraping protocol that respected rate limits and ethical data use standards.

5.3. Filtering and Preprocessing

Articles were filtered using a two-tier purposive strategy. First, content-based filtering identified only those articles explicitly discussing CEOs or executive teams in relation to strategic, financial, or organizational themes. Second, context-based filtering segmented firms into "innovative" or "traditional" sectors based on R&D intensity, technological positioning, and historical innovation performance.

Textual preprocessing included:

- Removal of boilerplate and duplicated news stories.
- Cleaning of metadata (source, publication date, industry tag).
- Normalization of text for NLP processing, including lemmatization and removal of HTML artifacts.

5.4. Construction of the CEO Overconfidence Index

Three sentiment analysis methods were employed to estimate overconfidence:

5.4.1. Conf(Press) Keyword Method

Adapted from Hirshleifer, Low, and Teoh (2012), this rule-based approach classifies press articles based on their tone when referring to CEOs. Specifically, each article is evaluated for the presence of keywords associated with confident or cautious language.

An article is labeled as reflecting overconfidence if confident terms outnumber cautious ones:

$$\text{Confident CEO (Press)}_i = \begin{cases} 1 & \text{if } \sum_{s=1}^{t} a_{is} > \sum_{s=1}^{t} b_{is} \\ 0 & \text{otherwise} \end{cases}$$

Where:

- a_{is} : indicator that article s, associated with CEO i, contains confident terms
- b_{is} : indicator that article s contains cautious terms
- t : total number of articles analyzed per CEO
- The output is a binary classification for each article referring to CEO i

Each article is treated as an independent unit, allowing classification at the document level. This contrasts with aggregate, firm-year measures used in prior literature but retains their core methodological logic. The keyword sets are applied uniformly across sectors and time periods to ensure consistency.

By working at the article level, this implementation enables finer-grained analysis of overconfidence dynamics across firms, events, and media outlets, and allows integration into downstream natural language classification tasks.

5.4.2. VADER Polarity Analysis

Using the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon, each article was assigned four sentiment scores: positive, neutral, negative, and compound. The compound score (–1 to +1) was used to evaluate overall tone. Articles with compound scores above 0.75 were flagged as likely overconfident and compared to the Conf(Press) labels to assess agreement. This method introduced a continuous perspective into what is otherwise a binary framework.

5.4.3. GenAI Classification

Using Google's Gemini 1.0 Pro model, full-text articles were fed to the LLM to obtain a confidence classification: *Confident* or *Unconfident*. The prompt included background on overconfidence framing and examples of both confident and cautious language, but the model was not restricted to these keywords. It evaluated semantic context, tone, and rhetorical structure to make a classification decision.

Each result included a brief rationale (i.e., justification text) for transparency and traceability. Ambiguous cases—such as contradictory language, sarcasm, or complex attributions—were manually reviewed to better understand model behavior at the edge.

The exact prompt template used for classification is provided in Annex C.

5.5. Aggregation and Index Construction

Each sentiment classification method produced binary outputs per article: either *Confident* (1) or *Unconfident* (0). These were then aggregated to create firm-level indices of CEO overconfidence.

For each firm *f* and method *m*, the overconfidence index is defined as:

$$\text{CEO OC Index}_{f,m} = \frac{\sum_{i=1}^{n_f} OC_{i,m}}{n_f}$$

Where:

- $OC_{i,m}$: indicator equal to 1 if article *i*, associated with firm *f*, was classified as overconfident by method *m*; 0 otherwise
- n_f : total number of articles available for firm *f*
- *f* : firm identifier
- *m* : classification method used (Conf(Press), VADER, GenAI)

This yields the proportion of a firm's articles classified as overconfident, enabling comparison across methods and sectors. The classification labels were generated in Python, and aggregation was performed in Tableau using consistent binary logic across methods.

5.6. Data Validation and Reliability

Several steps were taken to enhance reliability:

- Concordance: Results from GenAI, Conf(Press), and VADER were compared.
- Manual Checks: A sample of GenAI-classified articles were reviewed by the researcher.
- Filtering: Articles with low word counts or zero signal terms were removed.
- Balance: The final dataset included a near-even split between innovative and traditional firms (as visualized in Chapter 8).

In addition, articles flagged as having neutral or contradictory sentiment were excluded from the aggregated overconfidence index.

5.7. Ethical Considerations

All data were sourced through licensed platforms (ProQuest, EBSCO) or public APIs (New York Times) in compliance with terms of service. No private or sensitive information was accessed. AI classification outputs were limited to sentiment categories, and no executive identities were profiled or interpreted beyond their professional public portrayal.

6. Data Collection

This chapter details the process of sourcing, extracting, and preparing the dataset used for estimating CEO overconfidence. The data pipeline was built on three main pillars: ProQuest TDM Studio, EBSCO Information Services, and the New York Times Public API. These sources were selected to capture a wide array of CEO-related textual content from reputable publications across various time periods and industries. The process involved both automated and curated methods of retrieval, with a focus on transparency, breadth, and alignment with the study's analytical goals.

6.1. Data Sources

The corpus consists of 11,795 CEO-related articles drawn from three distinct platforms:

- ProQuest TDM Studio (used under a research trial): Provided access to premium business media including *The Economist*, *Fortune*, *Forbes*, *Financial Times*, and others.
- EBSCO's API: Accessed via university API credentials and filtered to include only full-text business journalism.
- New York Times Public API: Accessed via Python-based queries to retrieve articles relevant to executive leadership and firm decisions.

Despite differences in origin, format, and editorial style, all datasets underwent a structured curation and normalization process to ensure analytical consistency. This involved harmonizing metadata, filtering for full-text availability, and standardizing document formats across platforms. As a result, the same sentiment classification methods and criteria were applied uniformly to each article, regardless of source. This alignment allowed for valid cross-source comparisons and robust integration into downstream overconfidence analysis.

Table 6.1 summarizes the final contribution from each source.

Table 6.1: ProQuest TDM Studio Total Articles Extracted

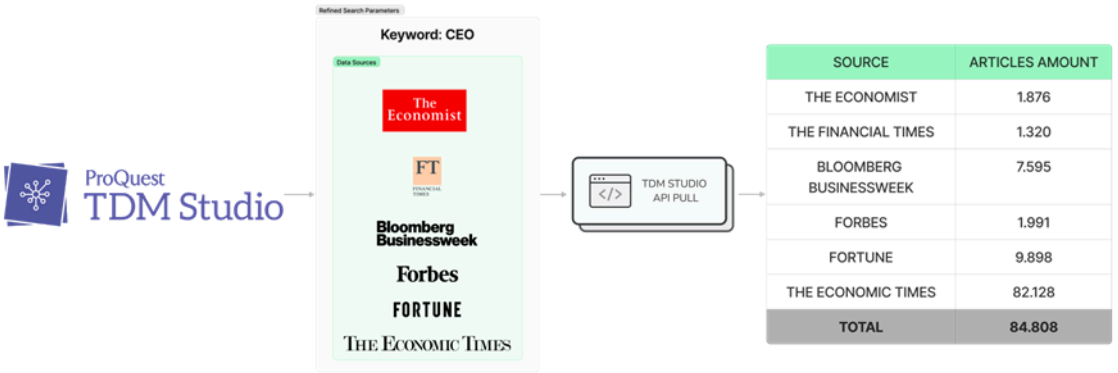
Source	Total Retrieved	Final Used
ProQuest (TDM Studio)	84,808	1,876
EBSCO API	33,758	7,919
NYT Public API	2,000	2,000
Total	—	**11,795**

6.2. ProQuest TDM Studio

The initial phase of data extraction centered on ProQuest TDM Studio, a platform designed for text mining from academic and commercial publications. Although initial queries returned over 3 million records for the keyword "CEO," access under the research trial was limited. To work within these constraints, the strategy was narrowed to focus on select high-reputation publications (e.g., *The Economist*, *The Financial Times*, *Bloomberg Businessweek*, *Forbes*, *Fortune*, *The Economic Times*).

Figure 6.1 presents an overview of the ETL process implemented for ProQuest data, from query design and source filtering to final extraction and cleaning. This visual highlights the decision points that shaped the final dataset structure and volume.

Figure 6.1: ProQuest TDM Studio ETL Overview



From the narrowed search scope, 84,808 documents were downloaded. However, due to limitations in the trial version—particularly in article volume limits and metadata quality—only 1,876 full-text articles from *The Economist* were retained. This subset was selected based on richness of content, clarity of sentiment, and publication consistency.

Table 6.2 details the number of articles retrieved per source, illustrating the dominance of *The Economic Times* in raw volume but also underscoring the richer content concentration found in *The Economist* subset.

Table 6.2: Articles Extracted from TDM Studio

Source	**Article Count**
The Economist	1,876
The Financial Times	1,320
Bloomberg Businessweek	7,595
Forbes	1,991
Fortune	9,898
The Economic Times	82,128
Total	**84,808**

TDM Studio posed challenges: high article volumes could not be processed efficiently, licensing limits prevented full access, and The Wall Street Journal or NYT content were unavailable through this platform. Nevertheless, *The Economist* subset offered the most coherent and sentiment-rich material for applying Conf(Press).

6.3. The New York Times API

To address the absence of NYT coverage in TDM Studio, a secondary pipeline was implemented using the NYT Public API. A script was developed to retrieve metadata and body content for articles containing the term "CEO". This pipeline included:

- Programmatic access via developer credentials
- Iterative querying with date and keyword filters
- JSON parsing to extract article abstracts, lead paragraphs, and publication dates

After cleaning and deduplication, 2,000 articles were selected based on clarity, length, and relevance to executive narratives. These articles were stored in structured format, lemmatized, and tokenized for later classification.

Table 6.3 summarizes the dataset construction parameters, including the date range (2009–2023), keyword filters, and exclusion criteria such as sports or opinion content.

Table 6.3: NYT Dataset Summary

Date Range	Articles Used	Filters Applied
2009–2023	2,000	Must mention CEO; exclude sports, obituaries, opinion

Although the NYT dataset was smaller, it introduced valuable temporal depth and stylistic variation. These articles were used for validation and testing of GenAI's ability to interpret different editorial tones.

6.4. EBSCO Information Services' API

The most technically challenging pipeline involved connecting to the EBSCO Discovery Service (EDS) API. After university-based authentication, the process involved:

- Creating sessions and managing access tokens
- Executing keyword searches and filtering results by source
- Using the "Retrieve" endpoint to access full-text articles
- Parsing results in JSON and transforming them into structured text

While initial pulls yielded over 33,000 results, only 7,919 full-text business articles were retained. Filtering criteria included source credibility (e.g., Bloomberg, WSJ, Fortune, Forbes), thematic focus, and length.

Table 6.4 summarizes the final article count by source. Each outlet contributed between ~2,000 and ~1,900 articles, ensuring a balanced distribution across top-tier business media.

Table 6.4: EBSCO Articles by Source

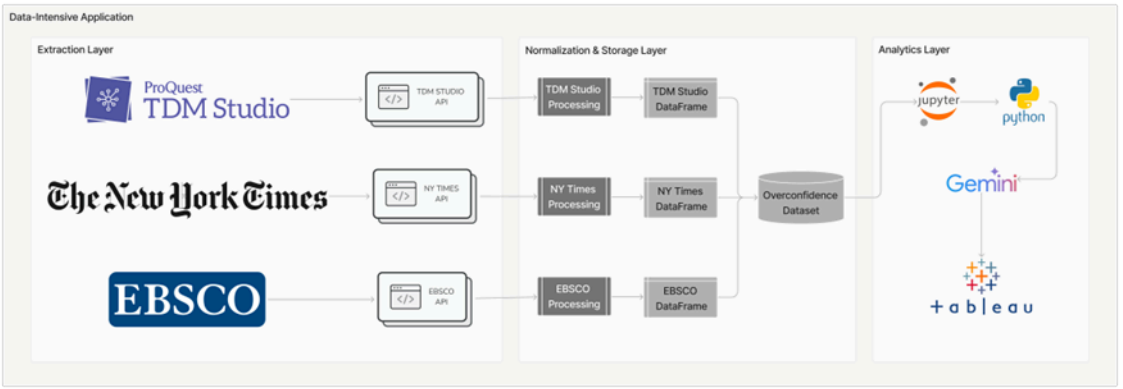
Source	Articles Used
Forbes	2,000
Fortune	2,000
Bloomberg	2,000
WSJ	1,919
Total	**7,919**

Technical difficulties encountered included limited filtering by publication, authentication expiration, and JSON structure inconsistencies. Nevertheless, the retained corpus was rich in narrative tone and structured metadata.

6.5. Data Integration and Infrastructure

Once extracted, all articles were processed into a centralized repository, enabling consistent data handling across sources and methods. The architecture supporting this workflow is illustrated in Figure 6.2, which presents the three-layer data infrastructure developed for this thesis.

Figure 6.2: Three-Layer Data Infrastructure



- Extraction Layer: Source-specific scrapers and API connectors used to ingest articles from ProQuest, EBSCO, and the NYT API
- Normalization Layer: Preprocessing steps such as deduplication, metadata standardization, and lemmatization to ensure uniform structure and language representation
- Analytics Layer: Application of sentiment analysis methods (Conf(Press), VADER, GenAI), followed by classification and sector-level segmentation

This modular pipeline allowed for scalable and repeatable analysis while preserving flexibility for integrating new datasets or testing alternative NLP models. By structuring the data flow into

discrete, reproducible layers, the infrastructure supports transparency and extensibility—key requirements for longitudinal sentiment analysis and behavioral finance applications.

6.6. Summary and Lessons Learned

The data collection process combined trial-based academic resources (TDM Studio), institutional access (EBSCO), and public datasets (NYT API). Each source presented unique trade-offs:

- TDM Studio offered volume, but limited by licensing
- EBSCO required heavy filtering and session-based access
- NYT API was simple but rate-limited and semantically diverse

The final integrated dataset consisted of 11,795 articles, aggregated from three distinct pipelines: 1,876 articles from *The Economist* via ProQuest TDM Studio, 7,919 articles from full-text EBSCO API results, and 2,000 articles from the New York Times Public API. These documents were selected based on a combination of textual richness, metadata availability, and direct relevance to CEO-related events or narratives.

While the initial retrieval across platforms yielded significantly higher volumes, extensive filtering was applied to ensure article quality, thematic alignment, and consistency in formatting. Articles with limited content, ambiguous attribution, or lacking executive-related context were excluded. The resulting corpus reflects a curated sample optimized for cross-source comparability and suitable for multi-method sentiment analysis. Data inclusion decisions were guided by legal access constraints, API limitations, and the overarching goal of maximizing interpretive value for overconfidence estimation.

Moreover, the dataset was designed to support both document-level and firm-level aggregation, enabling downstream comparisons across industries, companies, and time periods. This flexibility is essential for evaluating how CEO overconfidence manifests differently across contexts—a core aim of the analysis that follows.

These lessons informed the architecture of the CEO Overconfidence Index and its broader application, covered in the next chapter.

7. Exploratory Data Analysis (EDA)

This chapter presents the descriptive insights gained from analyzing the CEO-related article corpus prior to index construction. It serves both a diagnostic and conceptual purpose: to validate the suitability of sentiment-based methods, uncover sector-specific trends, and evaluate how sentiment manifests across press narratives. By combining lexicon-based scoring, keyword-driven classification, and GenAI-based interpretation, this exploratory phase confirms that the dataset is both semantically rich and methodologically tractable.

7.1. Dataset Integration and Preprocessing

The dataset comprised 11,795 articles obtained from three primary sources: ProQuest TDM Studio, EBSCO Information Services' API, and the New York Times Public API. Each source contributed distinct content types—firm-level profiles from ProQuest, broad sector coverage from EBSCO, and high-profile event narratives from the NYT.

After enforcing structural consistency and applying thematic filtering for executive relevance, 2,812 articles were retained for sentiment analysis. Preprocessing steps included:

- Removal of stopwords, punctuation, and URLs
- Lemmatization and lowercasing for consistent tokenization
- CEO-firm matching using Bloomberg tenure data and publication dates

A stratified manual review confirmed that retained articles reflected substantive discourse around CEO behavior or decision-making. Articles with only superficial executive mentions were excluded.

Challenges in preprocessing included ambiguous executive names (e.g., "Smith" or "Lee"), low-quality OCR outputs in older articles, and inconsistent naming conventions across time (e.g., "chief executive" vs. "CEO"). Additional preprocessing logic was implemented to handle edge cases, including fuzzy string matching and exclusion of opinion columns and obituaries.

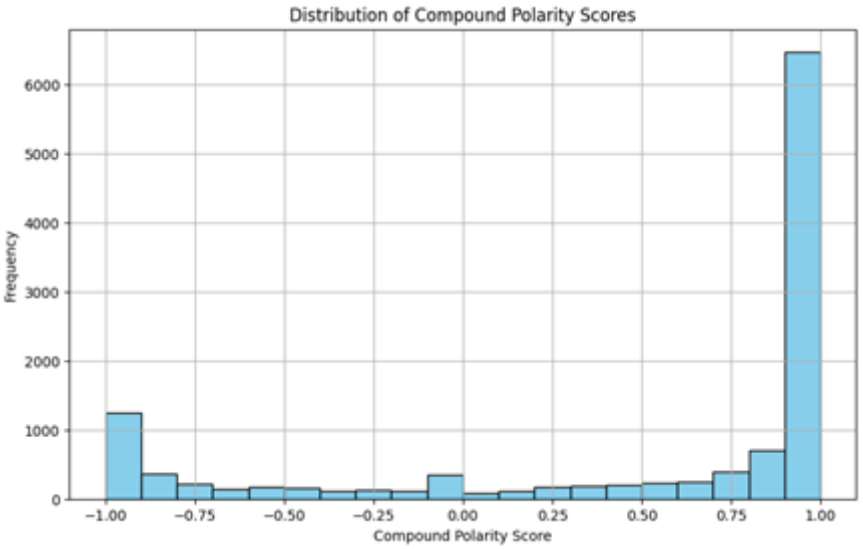
7.2. Sentiment and Polarity Distribution

Sentiment scoring employed VADER, which produces four scores per article: positive, negative, neutral, and compound. The compound score, ranging from -1 to +1, served as the primary polarity indicator for this analysis due to its aggregation of overall sentiment strength.

Across the dataset, the average compound score was 0.483, while the median score reached 0.936, indicating a pronounced right skew and suggesting an optimism bias in CEO media portrayal. A standard deviation of 0.722 further reflects high variability in sentiment across articles.

Figure 7.1 visualizes this distribution, showing that compound scores cluster heavily on the positive side. The drop-off toward negative values is steep, reinforcing the idea that critical or negative portrayals of CEOs are relatively rare in business journalism.

Figure 7.1: Distribution of VADER Compound Sentiment Scores



To illustrate this semantic skew, consider the following examples:

- *Positive polarity*: "Ford's new CEO praised for visionary realignment of electric strategy." → compound score > 0.9
- *Negative polarity*: "CEO criticized for strategic missteps and declining earnings." → compound score < –0.4

These findings are consistent with prior literature on media tone and executive coverage, reinforcing the need for context-aware models that go beyond lexical cues.

Figure 7.2: Word Cloud of Article Titles

Additionally, Figure 7.2 presents a word cloud of high-frequency terms in article titles. Common tokens like *"CEO," "company," "leadership,"* and *"investors"* confirm the business-centric framing of the dataset. Less frequent but semantically dense terms—such as *"turnaround," "visionary,"* and *"criticism"*—hint at narrative polarity and frame-setting around executive behavior.

7.3. Overconfidence Classification with Conf(Press)

The Conf(Press) method (Hribar and Yang, 2016) was used to classify overconfidence based on keyword frequency. Articles were labeled "Confident" if confidence-related terms (e.g., "optimistic," "confident," "visionary") outnumbered caution-related terms (e.g., "prudent," "steady," "conservative").

This analysis was applied to the 2,812 filtered articles. Articles with no relevant keywords were excluded, ensuring interpretive clarity.

Among the 1,342 articles retained in the final classification sample, 694 (51.7%) were labeled as "Confident" under the Conf(Press) method. This distribution highlights the relative conservatism of the approach and its dependency on explicit language.

Unlike polarity scoring, Conf(Press) focuses on targeted executive-related language. However, its primary limitation is its inability to detect confidence when conveyed through framing or implication.

Mini Case Example (Confident): *"Tesla Misses Big In Q1 As Its Stock Tanks"* — The article stated: *"CEO Elon Musk doubled down on autonomy and AI efforts, asserting Tesla's vision remains unchanged despite financial setbacks."* This was labeled Confident due to the assertive tone and strategic conviction expressed.

False Negative Example (Missed by Conf(Press)): *"The CEO moved forward with a controversial restructuring despite analyst skepticism."* While the decision demonstrates boldness, the absence of explicit keywords like "confident" or "visionary" led the method to classify it as Unconfident.

Approximately 36% of the total integrated articles were excluded from Conf(Press) scoring due to absence of any signal words, highlighting a key limitation in rigid keyword dependency.

7.4. CEO-Company Matching and Sector Tagging

To enable firm-level and sectoral analysis, each article was linked to a specific company using CEO tenure data sourced from Bloomberg, along with standardized ticker mappings. Industry classification was conducted using Standard Industrial Classification (SIC) codes, which provided a structured basis for assigning firms to broad sectors.

Following the framework proposed by Hirshleifer, Low, and Teoh (2012), each sector was then categorized as either innovative or non-innovative, based on historical citation-weighted patent output. This distinction enables sharper contrasts in later stages of the analysis, particularly in comparing how overconfidence manifests across industry types.

Table 7.1 presents the final sector classification schema. It lists representative SIC codes, corresponding sector descriptions, and the assigned innovation flag. For example, sectors such as *Technology*, *Healthcare*, and *Industrial Machinery* are classified as innovative, while *Food & Drink*, *Retail*, and *Insurance* are treated as non-innovative.

Table 7.1: SIC-Based Sector Classification and Innovation Mapping

SIC Code	Sector Description	Industry Classification	Innovation Flag
20	Food & Drink Products	Consumer Staples	Non-Innovative
28	Chemicals & Pharmaceuticals	Healthcare/Pharma	Innovative
35	Industrial Machinery & Computer Hardware	Manufacturing	Innovative
36	Electronic Equipment	Technology	Innovative
37	Transportation Equipment	Industrial Goods	Innovative
58	Retail – Eating & Drinking Places	Consumer Services	Non-Innovative
63	Insurance	Financials	Non-Innovative
73	Business Services (Eng., Accounting, Research)	Services	Innovative
87	Business Services – Research & Development	Services	Innovative

In addition to this classification, firm-level attributes—such as R&D expenditure—were also appended. These variables support contextual interpretation in later sections (e.g., Section 8.5) when analyzing the relationship between confidence levels and strategic corporate behavior.

7.5. Generative AI Classification Output

To address the limits of keyword-based sentiment, a Generative AI classification was implemented using Gemini 1.0 Pro. The model was prompted to classify each article as "Confident" or "Unconfident," based on narrative tone, executive portrayal, and strategic language.

Example Prompt Structure:

- *"Based on the article provided, does the CEO exhibit a confident or unconfident tone in their actions, language, or public statements? Evaluate based on strategic posture, leadership framing, and decision-making signals."*

Sample Model Response (Confident):

- *"The article describes the CEO of Tesla as 'doubling down on autonomy and AI,' emphasizing that their strategic vision remains unchanged despite short-term losses. This signals a bold and confident leadership style. Classification: Confident."*

Sample Model Response (Unconfident):

- *"The article highlights the CEO's cautious messaging following the company's weak earnings, stressing the need for measured decisions and minimizing risk. The tone is cautious rather than assertive. Classification: Unconfident."*

Rather than relying on keyword counts, the model interpreted broader semantic context, allowing for sentiment detection even in abstract or metaphorical language. This proved especially effective in articles with ambiguous tone or subtle rhetorical cues.

Out of the 1,114 labeled articles returned by the GenAI pipeline, 752 (67.5%) were categorized as "Confident," suggesting that the model captured confident portrayals in more nuanced or indirectly expressed formats than Conf(Press). An additional 228 articles yielded either null or ambiguous outputs and were excluded.

Table 7.2 presents a set of representative classification examples. Each row includes:

- A partial excerpt of the model's raw textual output
- The classification label assigned by the model (Confident or Unconfident)
- A short summary of the rationale, capturing the key reasoning behind the classification

For instance, the model identified assertive leadership framing in a case where the CEO was "doubling down on autonomy and AI," despite financial setbacks—something that rule-based methods would not have captured. In another case, the model classified an article as *Unconfident* because it emphasized caution, risk aversion, and lack of forward momentum in executive messaging.

These examples illustrate the GenAI model's ability to make context-sensitive decisions and surface non-explicit signals of overconfidence, offering a valuable complement to lexical and rule-based methods.

Table 7.2: Sample Results of Generative AI Application

GEN_AI_GEMINI_OUTPUT	GEN_AI_CEO_CONFIDENCE	GEN_AI_CEO_SUMMARY
Confident\n\nThe article discusses the tre...	Confident	The article discusses the trend of companies d...
Unconfident\n\nThe article does not mentio...	Unconfident	The article does not mention the CEO's confide...
Confident\n\nThe article portrays the CEO,...	Confident	The article portrays the CEO, Jensen Huang, as...
Confident\n\nThe article mentions that the...	Confident	The article mentions that the CEO, Jensen Huan...
Confident\n\nThe article is about the New ...	Confident	Overall, the article presents a positive and o...

7.6. Triangulating Sentiment: Why Multiple Methods?

Each sentiment method offered distinct strengths and weaknesses:

Method	Strengths	Limitations
VADER	Efficient, general polarity measure	Not specific to CEOs or confidence
Conf(Press)	Focused on confidence/caution framing	Rigid keyword dependency
GenAI	Context-aware, handles nuance	Lower transparency, higher cost

Triangulating across these methods provides both robustness and interpretability. Divergence between methods flags articles with nuanced tone or rhetorical framing, while convergence reinforces classification confidence. For example, articles labeled as "Confident" by both Conf(Press) and GenAI tended to feature bold strategic claims, leadership metaphors, or repeated investor praise.

7.7. Summary of the Exploratory Data Analysis

Preliminary analysis by sector reveals clear variation in how CEO sentiment is framed. Innovative industries—particularly those in technology, biotech, and electronics—tend to feature a higher proportion of articles portraying CEOs in confident terms. This may reflect greater investor expectations and a cultural emphasis on visionary leadership in these fields.

Example Insight: Articles from the software sector frequently invoke themes like "industry disruption" and "breakthrough thinking," portraying CEOs as bold agents of change. In contrast, coverage in financial services tends to highlight prudence and stability, with phrases such as "steady leadership" or "disciplined execution."

When classifications are cross-validated using GenAI, the divergence between sectors becomes more pronounced. The model flagged a significantly higher share of confident portrayals in innovative industries—even in cases where Conf(Press) failed to detect any relevant sentiment. This suggests that rhetorical style and media framing are highly context-dependent, and that traditional keyword-based methods may systematically underrepresent overconfidence in high-growth or high-risk sectors.

Key Takeaways:

- Sentiment expression varies significantly across sectors, shaped by context and rhetorical norms.
- GenAI is particularly effective at detecting implied confidence in sectors where strategic language is more nuanced or aspirational.
- Incorporating sector-sensitive sentiment methods improves the accuracy of behavioral trait detection, such as overconfidence.

These findings strengthen the rationale for a hybrid, multi-method CEO Overconfidence Index and lay the groundwork for the comparative analysis presented in the following chapter.

8. Findings and Data Product Insights

8.1. From Index to Insights: Building a Behavioral Data Product

This section presents the empirical insights derived from the CEO Overconfidence Index, produced through the integrated press- and GenAI-based pipeline introduced in earlier sections. The results are visualized using Tableau Software, which allows for dynamic interaction with the underlying data. Our findings are organized thematically, beginning with overconfidence segmentation by industry, followed by press activity analysis, and concluding with the relationship between overconfidence and R&D investment.

The CEO Overconfidence Index is deployed as a full-stack data product, integrating multiple data layers—textual sentiment, financial disclosures, and firm-level attributes—into an interactive tool for insight generation. Tableau dashboards were built to enable exploration across time, firm, and industry. This application supports two key goals:

1. Monitoring behavioral risk in near real-time, via dynamic dashboards and confidence scoring.
2. Benchmarking sectoral overconfidence patterns, enabling comparison across innovative and traditional firms.

Key components of the data product include:

- Interactive Dashboards: Visual interfaces that allow filtering by industry, firm, or CEO-level attributes.
- Automated Reports: Scheduled exports for analysts and governance teams.
- Custom Alerts and Predictive Analytics: Notification systems that trigger when thresholds are crossed.

The product is modular and scalable, supporting integration of new data sources such as earnings calls, regulatory filings, and financial KPIs.

8.2. CEO Overconfidence Across Sectors

A core contribution of this study is the ability to distinguish overconfidence patterns across innovative versus non-innovative industries. Across the full dataset, 191 distinct CEOs were classified as *Confident* at least once, spanning a total of 32 years of coverage.

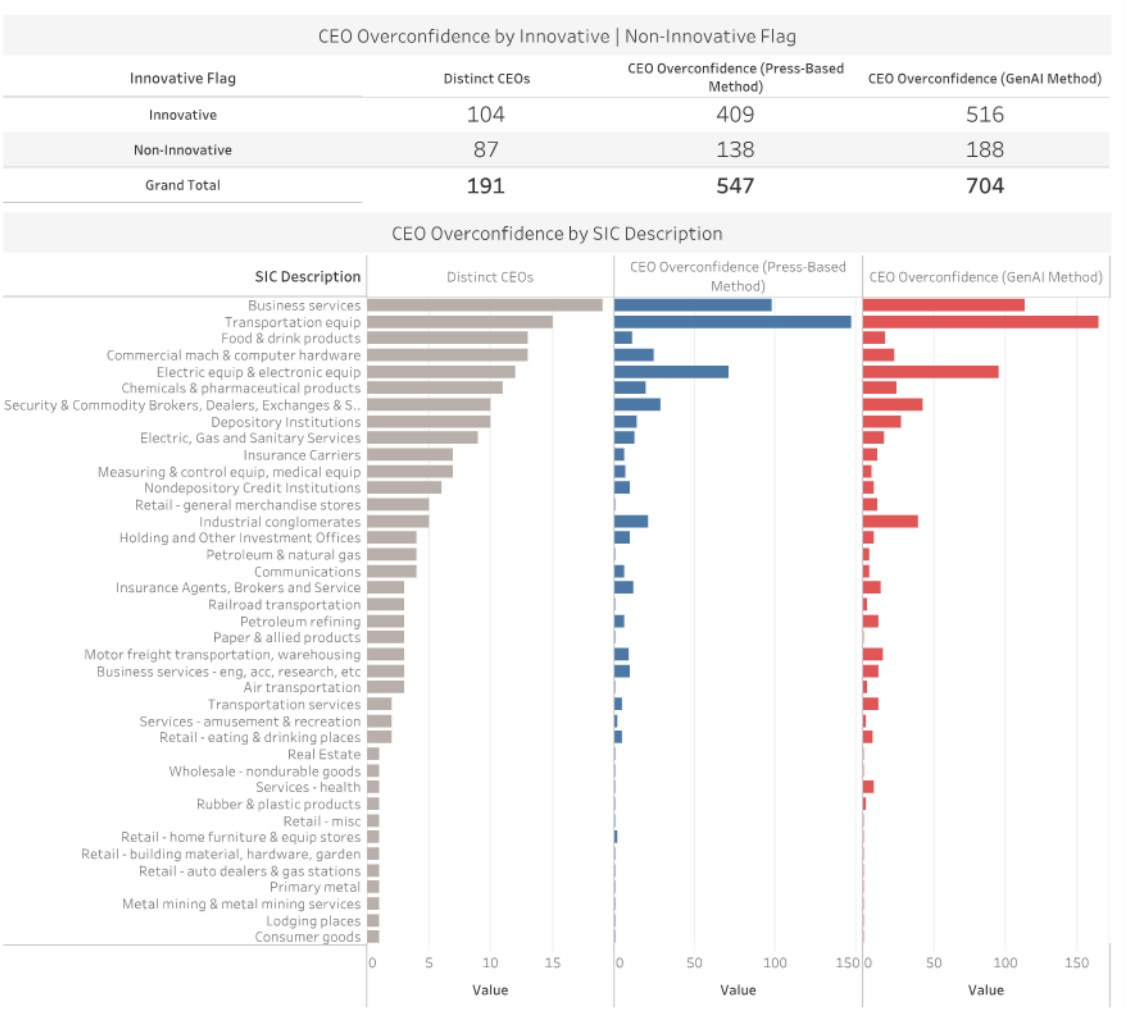
Table 8.1 summarizes the distribution of confident CEOs by sector type. Innovative sectors, such as *Business Services*, *Electronics*, and *Transportation Equipment*, account for a slightly larger share of distinct confident CEOs (104 out of 191). Non-innovative sectors include industries like *Food Products*, *Insurance*, and *Depository Institutions*, with 87 CEOs flagged.

To assess overconfidence intensity at the industry level, sentiment scores were aggregated by SIC code using two classification methods: Conf(Press) and GenAI. Figure 8.1 presents a dashboard summary of CEO overconfidence distributions across industries.

Table 8.1: Summary of Confident CEOs by Sector

Sector Type	Distinct CEOs	Years Covered	Notable Industries
Innovative	104	28	Business Services, Electronics, Transportation Equipment
Non-Innovative	87	29	Depository Institutions, Food Products, Insurance

Figure 8.1: CEO Overconfidence Summary Dashboard



This visualization reveals clear clustering of overconfidence in innovative sectors, particularly in fields like *Electronics* and *Business Services*. Additionally, the dashboard shows a systematic difference in detection rates between methods:

- GenAI identified 516 overconfident articles for innovative firms.
- Conf(Press) identified 409, a lower count using stricter keyword-based thresholds.

This discrepancy underscores GenAI's advantage in interpreting implicit or contextually embedded signals of overconfidence, which may be missed by rule-based approaches. It reinforces the value of applying multiple methods in triangulation, particularly when assessing behavioral traits in narrative-heavy sectors.

8.3. Comparative Confidence Across Firms

Figure 8.2: CEO Overconfidence Scatter Plot by Company

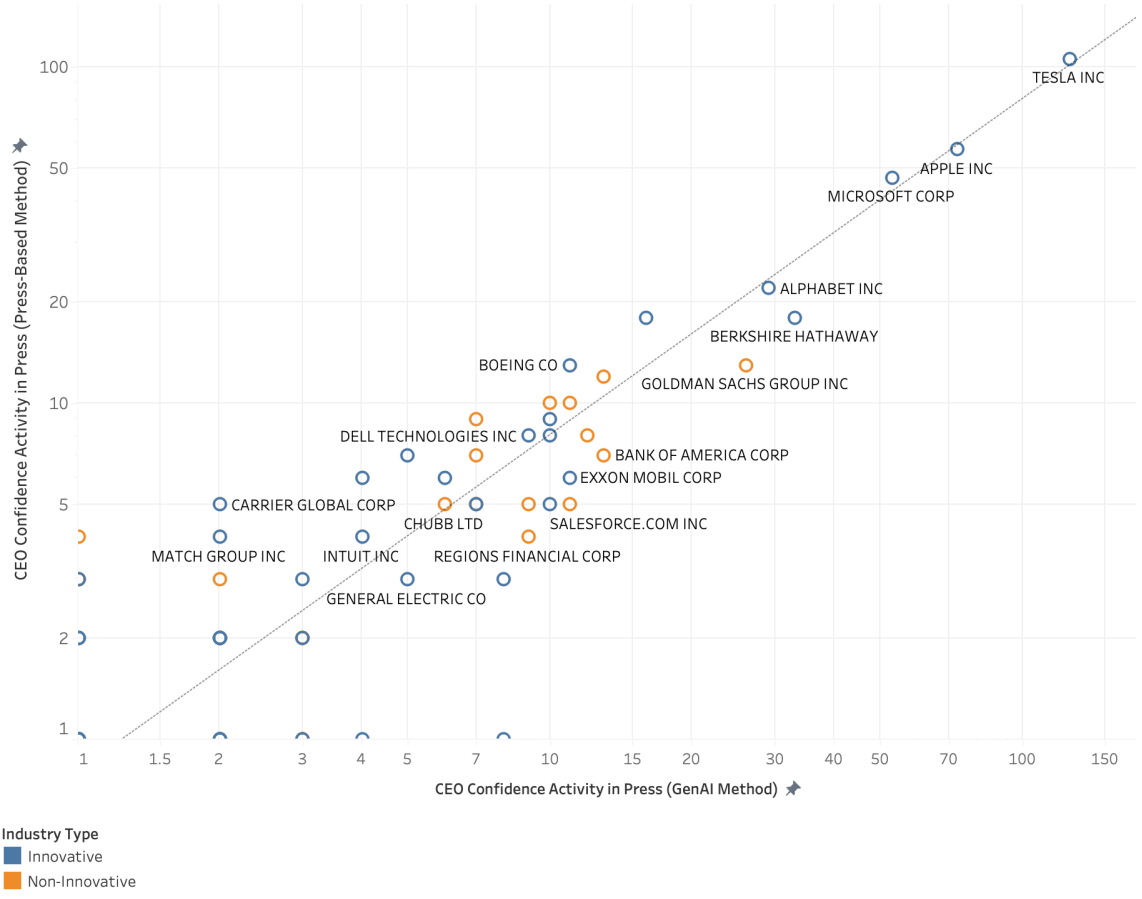


Figure 8.2 presents a scatterplot comparing firm-level overconfidence scores using two sentiment classification methods: Conf(Press) and GenAI. Each point represents a company, with color indicating whether the firm is classified as innovation-intensive (blue) or traditional (orange). The dashed 45° reference line denotes parity between methods; points near the line indicate agreement, while greater vertical or horizontal deviation reflects divergence in classification.

Most innovation-driven firms—such as Tesla, Apple, and Microsoft—cluster near or above the diagonal, suggesting that both methods consistently detect assertive sentiment in their CEO-related press coverage. Tesla, in particular, registers the highest scores across both methods, implying a persistently strong tone of executive confidence.

In contrast, Goldman Sachs lies significantly below the diagonal: its score under GenAI is substantially higher than under Conf(Press). This indicates that GenAI captures rhetorical or

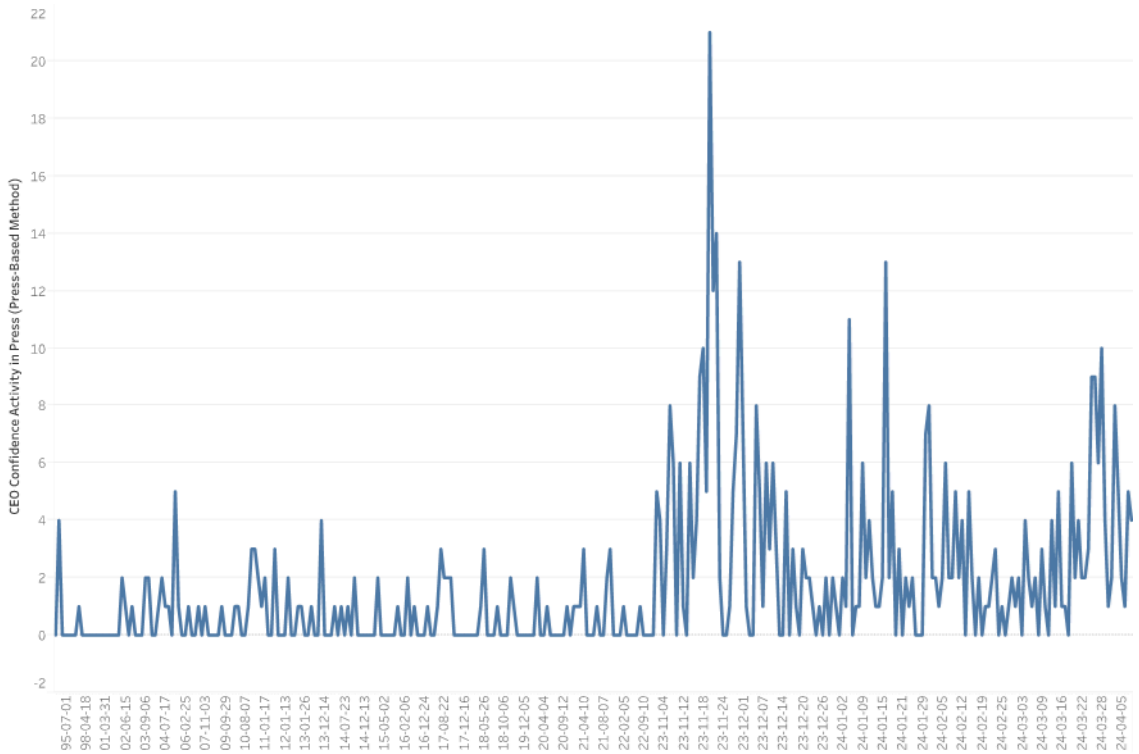
contextual indicators of overconfidence—such as personal framing, future-oriented claims, or semantic assertiveness—not well detected by lexicon-based methods.

These patterns highlight the value of triangulating sentiment methods. Methodological divergence can signal meaningful differences in linguistic framing, while convergence suggests stylistic uniformity or tone that is robust across detection models. The 45° reference line thus acts not only as a visual baseline but also as a diagnostic tool for interpreting how overconfidence manifests differently across firms and sectors.

8.4. Press Activity Over Time

To track how CEO overconfidence has evolved, sentiment classifications were aggregated over time using both the Conf(Press) and GenAI methods. Figure 8.3 displays the Net CEO Confidence Impact, defined as the difference between the number of *Confident* and *Unconfident* articles per period using the Conf(Press) method.

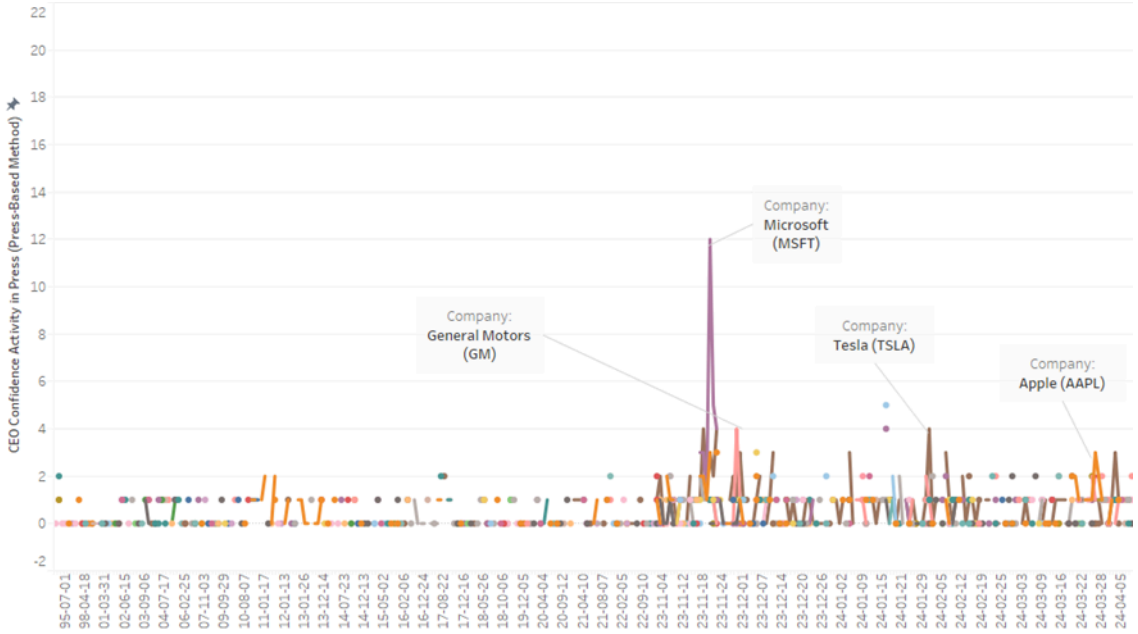
Figure 8.3: Net CEO Confidence Impact in Press Activity



Between 1995 and 2024, confidence levels show clear fluctuations, with notable peaks around landmark events such as the dot-com boom, the 2008 financial crisis, and the post-COVID surge in AI-driven narratives. Rather than a linear trend, the data reflects event-driven sentiment volatility, with confidence often spiking in response to moments of technological or strategic disruption.

To explore how CEO overconfidence varies by company, Figure 8.4 disaggregates net confidence scores by firm over time. The chart highlights individual trajectories for each firm, with particularly prominent confidence spikes for Microsoft (MSFT), Tesla (TSLA), and Apple (AAPL). These peaks align with high-visibility corporate events:

Figure 8.4: Net CEO Confidence Impact in Press Activity by Company

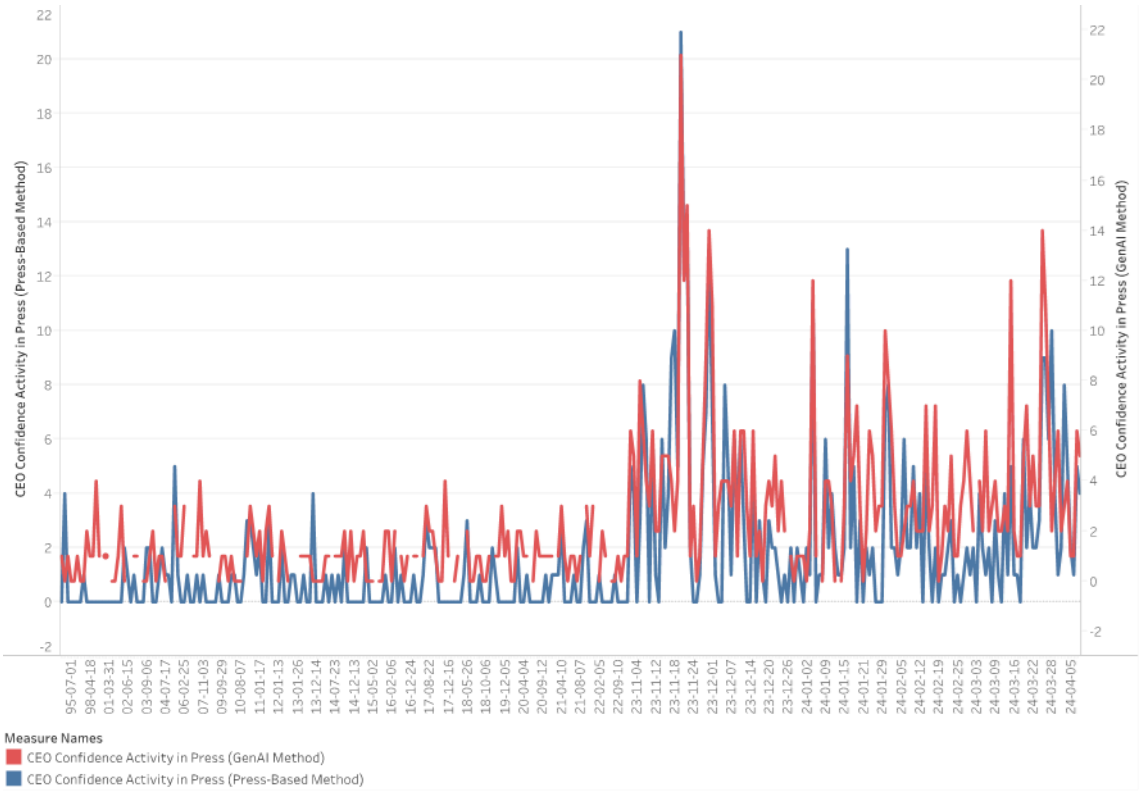


- Microsoft: Spikes in 2023 correspond to major AI-related announcements, such as its integration of OpenAI tools across enterprise products.
- Tesla: Recurring surges reflect Elon Musk's assertive communication style, especially around autonomy, pricing strategy, and investor communications.

The variation in confidence impact across firms illustrates how CEO communication style, media coverage intensity, and strategic positioning shape public sentiment over time.

Figure 8.5 compares the Net CEO Confidence Impact measured using Conf(Press) (blue) and GenAI (red). Both methods generally align during major sentiment events—such as the early 2023 confidence spike—but diverge in granularity and responsiveness.

Figure 8.5: Net CEO Confidence Impact in Press Activity (Press-based v. GenAI)



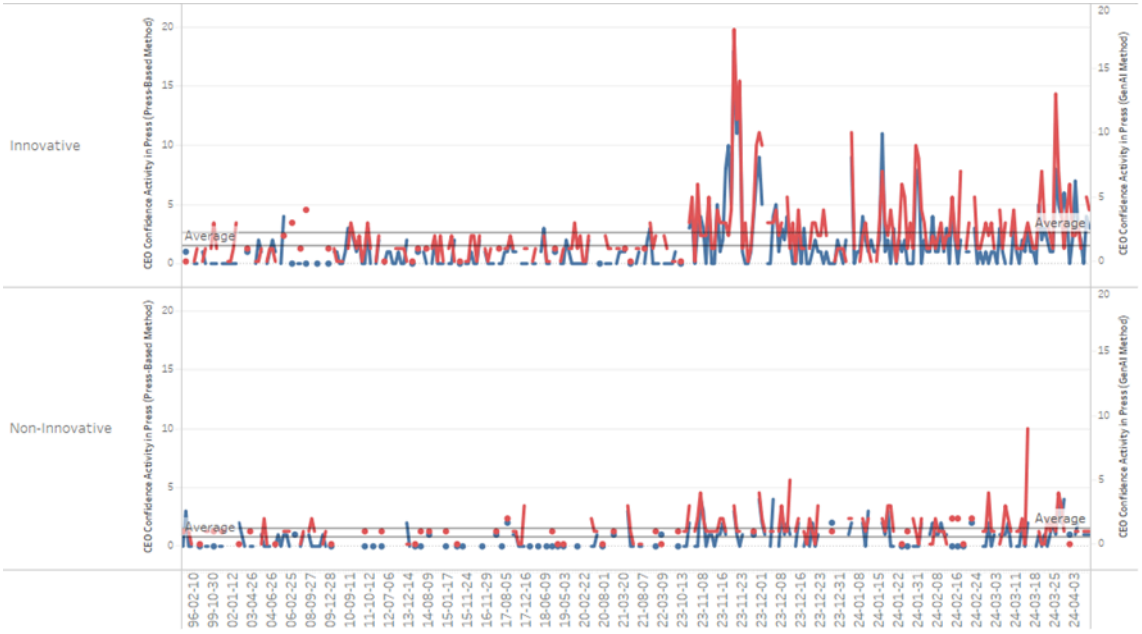
Post-2020, GenAI consistently detects more frequent and pronounced fluctuations, especially in periods with ambiguous tone, complex narratives, or emerging trends (e.g., AI ethics, corporate governance, tech sector regulation). The higher sensitivity of GenAI in this period reflects its strength in capturing contextual cues, such as indirect language or rhetorical framing.

This reinforces a central methodological insight: keyword-based methods like Conf(Press) provide a useful baseline but may under-detect nuanced shifts in executive tone. In contrast, GenAI offers a more adaptive lens for behavioral analysis, particularly in media environments where confidence is communicated implicitly.

8.5. Confidence and Innovation: The Sectoral Lens

To explore how innovation intensity influences CEO sentiment, Figure 8.6 compares net CEO confidence activity across innovative and non-innovative sectors, as classified via SIC codes. The time-series plots show confidence impact over time, using both Conf(Press) (blue) and GenAI (red) methods.

Figure 8.6: Net CEO Confidence Impact in Press Activity by Innovative Flag



In the upper panel, innovative firms exhibit greater sentiment volatility, especially after 2012. Confidence spikes in this group align with key industry milestones—such as the release of ChatGPT, advances in autonomous vehicles (e.g., Tesla FSD), and major AI investments by Big Tech. GenAI registers more pronounced peaks than the press-based method, reflecting its sensitivity to implicit confidence cues in speculative or forward-looking narratives.

In contrast, the lower panel shows that non-innovative firms maintain flatter sentiment trends, with fewer and smaller spikes. Occasional peaks occur in response to sector-specific events—such as post-COVID recovery strategies in banking and pharma—but the overall pattern suggests less dramatic shifts in perceived CEO overconfidence.

These results reinforce the theoretical expectation that innovation-intensive sectors are more exposed to narrative amplification of confidence traits. In industries where risk-taking, vision-setting, and public storytelling are central to strategic positioning, confidence tends to surface more frequently—and more variably—in media discourse.

8.6. Overconfidence and R&D Investment

Figure 8.7: R&D Expenses Evolution for Confident CEOs Companies

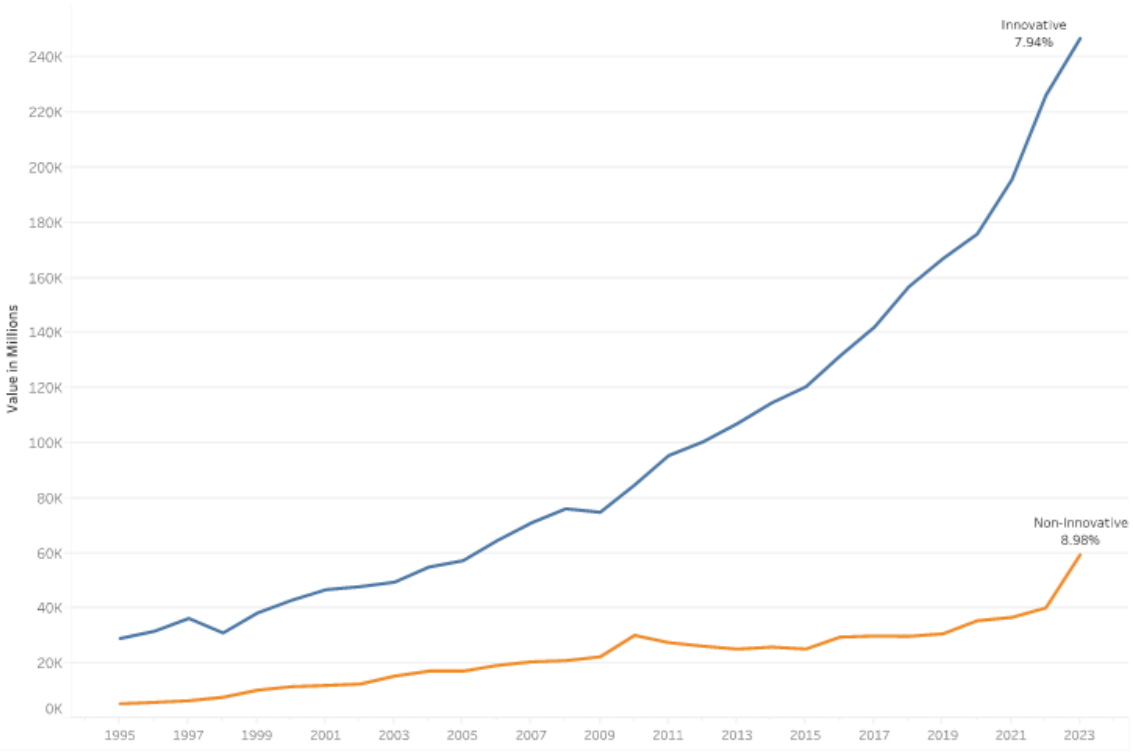


Figure 8.7 examines the relationship between CEO overconfidence and firm-level investment in research and development (R&D). Results suggest a meaningful link, particularly within innovation-driven sectors.

- Firms in innovative industries invest 4.5 times more in R&D compared to traditional sectors.
- While non-innovative sectors exhibit a slightly higher growth rate (8.98% CAGR), this is measured from a lower base of R&D spending.
- The reported percentages represent the compound annual growth rate (CAGR) of R&D expenses over the analysis period, calculated for each industry group based on firm-level expenditure data.

These findings suggest that high baseline R&D intensity may reinforce CEO overconfidence, particularly in fast-evolving sectors where technological bets are strategic and high-stakes. Notably:

- CEOs with sustained high confidence scores are often found in firms within the top R&D expenditure quartile.
- Fluctuations in confidence scores appear more correlated with industry growth momentum than with firm size or market capitalization.

8.7. Robustness and Triangulation

To validate the reliability of the CEO Overconfidence Index, we examined:

- Temporal coherence: Confidence scores rise in known optimism cycles (e.g., dot-com boom, AI wave).
- Cross-method convergence: Most highly confident CEOs (e.g., Musk, Nadella) score highly in both methods.
- Discrepancy analysis: When methods diverge, GenAI often captures subtle tone missed by lexical scoring.

This triangulation confirms both internal consistency and methodological complementarity.

8.8. Summary

This section demonstrates how an integrated, sentiment-based data product reveals behavioral dynamics across firms, sectors, and time. Key takeaways include:

- Overconfidence is more prevalent in innovative sectors.
- GenAI classification is more sensitive to context and implicit language than static press methods.
- Confidence spikes align with major technological or leadership events.
- R&D intensity may act as both a cause and consequence of overconfidence.

Together, these findings support the value of a dynamic, multi-source Overconfidence Index as a governance and decision-making tool.

9. Discussion

Interpreting the results within the scope of the research questions reveals a differentiated relationship between CEO overconfidence and industry context. The data shows a consistent pattern: companies operating in innovation-driven sectors tend to exhibit higher levels of CEO overconfidence. This pattern is especially pronounced among firms like Tesla, Apple, Alphabet, and Microsoft, aligning with the theoretical expectation that environments characterized by uncertainty, rapid change, and outsized rewards incentivize bold executive behavior. The findings provide empirical support to the proposition that executive sentiment varies across industry sectors and suggest that overconfidence may, in certain contexts, play a functional role in driving innovation.

This interpretation builds on the behavioral finance literature which posits that overconfidence can lead to both productive risk-taking and costly errors (Malmendier and Tate, 2005; Hirshleifer et al., 2012). In innovative industries, the potential upside of visionary decisions often outweighs the risks of occasional miscalibration. Our results suggest that CEO overconfidence may be an endogenous response to the strategic imperatives of high-velocity environments. This does not imply that overconfidence is uniformly beneficial; rather, it suggests that the distribution of outcomes linked to CEO traits is shaped by the structural characteristics of the firm's industry.

The study also reinforces the role of media and industry narratives in influencing perceptions of leadership. Press coverage that emphasizes boldness, disruption, or visionary framing can reinforce or amplify overconfidence tendencies. This aligns with prior research highlighting the role of public attention and reputational feedback in shaping CEO behavior (Malmendier and Tate, 2008). Importantly, this mechanism may also introduce self-reinforcing biases, as overconfident behavior is valorized and subsequently emulated or rewarded.

In methodological terms, the comparison between traditional press-based sentiment analysis and the GenAI-driven classification framework adds a second layer of insight. The traditional approach, exemplified by Conf(Press), offers historical depth and interpretability, allowing researchers to identify overconfidence through a pre-defined set of keywords in media coverage. However, this method can suffer from variability in journalistic language, limited ability to detect nuance, and exposure bias due to uneven media attention.

The GenAI approach, in contrast, leverages large language models to analyze text in context. This allows for more sophisticated recognition of implicit confidence expressions, semantic patterns, and changes in tone over time. Unlike keyword-based approaches, GenAI classification adapts to phrasing variation and can interpret overconfidence within broader narrative structures. Thus, GenAI enables fine-grained differentiation across large datasets, potentially capturing temporal shifts, implicit language cues, and multi-source corroboration. The integration of these methods enhances the robustness of the findings and opens new possibilities for monitoring CEO sentiment dynamically.

The inclusion of a third method, VADER (Valence Aware Dictionary and sEntiment Reasoner), further contextualizes these findings. As a lexicon- and rule-based sentiment analysis tool optimized for social and short-form text, VADER offers a midpoint between keyword approaches and GenAI. While it provides a normalized sentiment score and handles intensifiers or negation better than simple keyword counts, its fixed vocabulary limits adaptability across more nuanced contexts. In this study, VADER served as a valuable benchmark, particularly for sentiment polarity, but its limitations in capturing implicit confidence reinforce the importance of combining it with both

GenAI and traditional methods. This triangulation of sentiment estimation—using Conf(Press), VADER, and GenAI—enhances methodological rigor and allows cross-validation of CEO sentiment patterns from multiple analytic lenses.

From a data product perspective, the CEO Overconfidence Index developed herein offers actionable insights. It can serve as a monitoring tool for governance bodies, investors, and analysts to flag sentiment shifts, contextualize executive decisions, or assess firm-level leadership risk. As Galasso and Simcoe (2011) argue, recognizing the behavioral attributes of CEOs is essential for informed oversight, especially in sectors where risk-taking is tightly coupled with value creation.

While the findings are promising, they must be interpreted within the limits of the data and methodology. Nevertheless, the alignment between empirical patterns, theoretical expectations, and method triangulation strengthens the validity of the observed relationships. The study's contribution lies not only in the evidence it presents but also in the methodological architecture it proposes for future research into behavioral traits in executive decision-making.

To close the loop with the core research questions, Table 9.1 outlines how each was addressed:

Table 9.1. Summary of Research Questions and Corresponding Findings

Research Question	Findings & Interpretation
How does CEO overconfidence vary across industries?	The analysis shows higher levels of CEO overconfidence in innovation-driven sectors, suggesting that industry dynamics shape executive sentiment.
How do traditional and advanced sentiment methods compare in detecting overconfidence?	GenAI outperforms Conf(Press) and VADER in nuance and contextual understanding, while triangulation enhances reliability.
Can overconfidence be monitored through a data product for practical use?	The CEO Overconfidence Index provides a scalable and actionable tool for governance and investment contexts.

10. Implications

This research has practical implications for multiple stakeholder groups. For corporate boards and governance bodies, the CEO Overconfidence Index offers a tool to monitor executive sentiment beyond financial indicators. By identifying patterns of overconfidence—particularly when such sentiment deviates from industry baselines—boards can intervene preemptively to recalibrate strategic direction or adjust executive incentives.

For investors, the ability to contextualize CEO sentiment by industry provides an additional layer of interpretive power. Overconfidence, while often viewed as a liability, may signal assertive leadership in sectors where such traits align with innovation-driven value creation. By incorporating CEO sentiment scores into due diligence or risk models, investors can better anticipate volatility or pivot points in executive strategy.

At the strategic level, firms operating in high-risk environments may benefit from frameworks that temper executive confidence through structured decision protocols or feedback mechanisms. Industry professionals, especially in sectors with high failure rates (e.g., biotech, early-stage tech), can use these insights to refine leadership development or succession planning models. These insights may also inform the design of risk oversight committees, scenario-planning exercises, or adaptive decision-making frameworks that seek to mitigate the volatility that often accompanies overconfident leadership.

For human capital and organizational development practitioners, sentiment-based indices can support performance evaluations, leadership coaching, and executive development pipelines. Tracking overconfidence over time could provide early warning signals of executive drift, misalignment with firm culture, or blind spots in leadership perception.

Researchers can also draw on this study's integration of traditional sentiment measures with GenAI-enhanced classification to explore other executive traits—such as resilience, risk aversion, or adaptability—at scale. This methodological synthesis contributes to advancing behavioral finance and opens the door for interdisciplinary collaboration with data science and computational linguistics.

Lastly, the development of a data product as an output reinforces the applied value of academic work. The ability to generate real-time or periodic assessments of CEO overconfidence offers practitioners a bridge between research and operational decision-making, aligning with the increasing demand for actionable behavioral metrics in corporate strategy.

Illustrative Case: WeWork and CEO Overconfidence

WeWork's meteoric rise and rapid collapse between 2016 and 2023 offers a textbook example of founder-led overconfidence in action. At its peak, the company reached a private valuation of $47 billion, largely propelled by charismatic CEO Adam Neumann, who championed a vision of "elevating the world's consciousness" through coworking spaces (Pendergraft, 2022). Major investments from SoftBank and an aggressive expansion strategy lent credibility to this narrative, even as the company accumulated billions in operating losses (Westbrook, 2020).

Several behaviors exhibited by Neumann align with the overconfidence signals tracked by this study's index: extreme risk-taking, promotional language with spiritual or grandiose overtones, and corporate governance lapses. These included related-party transactions (personally leasing

properties to WeWork), the use of dual-class shares to consolidate voting control, and the introduction of opaque performance metrics like "community-adjusted EBITDA" (Westbrook, 2020; Pendergraft, 2022).

In 2019, WeWork's S-1 filing for a proposed IPO revealed staggering financial losses and governance irregularities, prompting investor backlash. The IPO was withdrawn, Neumann was ousted from the CEO role, and a valuation collapse ensued—dropping from $47 billion to below $10 billion in less than a year. By November 2023, the company filed for Chapter 11 bankruptcy protection in the U.S. (Westbrook, 2020).

Table 10.1 below charts this trajectory, overlaying valuation milestones and critical governance events. While this pipeline was not live during the WeWork episode, it is plausible that its application to Neumann's public statements, media exposure, and linguistic sentiment patterns would have flagged substantial deviation from industry baselines, particularly in the final year leading to the IPO attempt.

This case illustrates the type of retrospective and interpretive analysis that the CEO Overconfidence Index is designed to support. While it is not intended as a predictive model in the strict sense, future work might explore backtesting the index against similar founder-led growth firms to better understand potential early warning signals and their practical interpretation.

Note on scope and ethical considerations: This example is presented solely as an illustrative, retrospective case to demonstrate how the index logic could be applied to a well-documented real-world scenario. No actual OC-Index values are calculated for Adam Neumann, as WeWork does not belong to the S&P 500 universe that defines the study sample. Furthermore, for ethical and confidentiality reasons, this thesis avoids presenting individualized results with the names of real CEOs from the actual dataset. The main analysis remains aggregated or sector-level to prevent personal labeling or unintended interpretation of specific executive behavior.

Table 10.1 – WeWork Valuation and Governance Events, 2016–2023

Year	Valuation Estimate (USD Bn)	Notable Events
2016	16.9	SoftBank begins major investment round
2017	20.0	Neumann purchases properties leased to WeWork
2018	42.0	Expansion accelerates; dual-class share structure solidified
2019	47.0	IPO filed; S-1 reveals governance concerns; IPO withdrawn; Neumann resigns
2020	9.0	SoftBank bailout completed; layoffs begin
2021	8.0	Merger with SPAC (BowX); trades under new valuation basis
2023	<0.1	Files for Chapter 11 bankruptcy (Nov)

11. Limitations

This research faces several limitations, grouped into three interrelated domains: data access and representativeness, methodological design, and conceptual framing.

Data Limitations

As Clive Humby famously stated, "Data is the new oil"—a concept later popularized by Teradata thought leadership—emphasizing the critical role data plays in deriving value through analysis. However, Access to data-rich platforms such as ProQuest TDM Studio or Twitter's API posed financial constraints that restricted the scope and diversity of data sources. Although trial access offered temporary relief, the reliance on a limited number of outlets introduces potential sampling bias. This may limit generalizability, particularly for firms with low media exposure or in geographies underrepresented in mainstream English-language outlets. Future research would benefit from open-access datasets, partnerships with data providers, or multilingual corpora to enhance coverage.

Methodological Constraints

Each sentiment analysis method used has intrinsic weaknesses. Keyword-based models like Conf(Press) lack semantic sensitivity and are prone to misclassification due to contextual nuances such as sarcasm or negation. VADER, while more linguistically attuned, is bound by a fixed lexicon and struggles with domain-specific jargon. GenAI, although highly adaptive, presents challenges around interpretability, prompt sensitivity, and computational resource demands. Moreover, the ensemble use of methods may introduce inconsistency in sentiment scale or classification thresholds. Balancing interpretability with accuracy remains a key methodological challenge.

Conceptual Ambiguity

Overconfidence remains a latent psychological construct. While this study approximates it through linguistic cues—such as expressions of assertiveness or optimism—these proxies may not fully capture internal states or cognitive miscalibration. Furthermore, language usage may reflect strategic communication choices or public relations considerations rather than genuine overconfidence. The inability to validate textual sentiment against internal decision-making or outcomes introduces uncertainty around construct validity.

Despite these limitations, the triangulated approach and explicit methodological reflection strengthen the study's credibility. By acknowledging constraints in data, method, and interpretation, the research lays a transparent foundation for future refinement and validation.

12. Summary

This study explores the presence and implications of CEO overconfidence across industries, particularly in innovation-driven sectors. Through a hybrid sentiment classification pipeline—combining keyword-based analysis (Conf(Press)), VADER scoring, and GenAI-powered contextual classification—the research identifies a consistent pattern: firms operating in high-innovation environments display more assertive and confident CEO communication profiles.

Central to this thesis is the development of the CEO Overconfidence Index, a comparative metric constructed from large-scale press text analysis. By linking sentiment classifications with industry context, the index provides a structured way to interpret executive behavior. The findings reinforce core theories in behavioral finance by showing that overconfidence is not uniformly distributed, but conditioned by the strategic demands and reputational dynamics of specific industries. Moreover, the findings point toward an important reframing of CEO behavior: overconfidence in certain environments may not be merely a bias but a strategic expression, reflecting the executive's adaptation to market expectations and innovation-driven leadership archetypes.

Methodologically, this work contributes a modular and scalable framework that aligns behavioral theory with natural language processing techniques. The GenAI component expands the analytical reach of sentiment models by introducing contextual nuance, overcoming some of the semantic limitations found in traditional or lexicon-based tools. This not only improves precision in classification but also sets the foundation for dynamic monitoring applications. VADER, in turn, serves as a middle-ground benchmark for polarity detection, and its role in validating sentiment direction further supports the value of multi-method analysis.

Practically, the data product supports governance, risk oversight, and investor strategy by offering insights into executive tone and decision-making orientation. Its longitudinal potential makes it suitable not just for snapshot evaluations but for tracking behavioral change over time. These capabilities are particularly relevant in today's fast-evolving leadership environments, where digital footprints, market narratives, and public perceptions continuously shape executive influence.

While challenges remain—particularly in data availability, model calibration, and interpretability—the framework demonstrates a proof of concept for the operational use of behavioral analytics in corporate contexts. The study also highlights critical areas for refinement, such as addressing linguistic variation across international firms and evaluating sentiment thresholds for actionable interpretation.

Ultimately, the thesis bridges conceptual and technical domains, showing how AI-enabled textual analysis can enhance our understanding of executive psychology and corporate behavior. It opens up avenues for methodological replication, cross-sector adaptation, and deeper empirical engagement with the behavioral dimensions of leadership. As industries grow increasingly data-driven and reputationally sensitive, the strategic interpretation of sentiment will become a vital lens through which executive conduct is understood, measured, and managed.

13. Future Research Directions

Sectoral Differentiation

Expand the scope of industry comparisons to examine overconfidence in capital-intensive vs. service-oriented firms, or in regulated vs. deregulated industries. Future research could also analyze sector-specific phenomena in real time. For instance, current trends in the tech sector—such as strategic pivots, high-risk investment announcements, and bold public communications—could serve as contemporary examples of how executive sentiment interacts with innovation cycles and market narratives.

Longitudinal Analysis

Track sentiment scores over CEO tenures to evaluate how overconfidence evolves with experience, firm performance, or market cycles. Emerging data from rapidly scaling tech companies and CEO transitions at firms like OpenAI or Meta could offer natural experiments.

Cross-Cultural Studies

Investigate whether cultural dimensions (e.g., individualism, uncertainty avoidance) influence how overconfidence is expressed and perceived. Comparing U.S.-based tech giants to Asian or European firms like Tencent, Samsung, or SAP may reveal structural differences in leadership expression and media portrayal.

Governance Moderation

Study how board composition, ownership structure, or institutional investor presence moderates the impact of overconfidence on firm decisions. Examining how firms like Apple or Amazon balance CEO prominence with board accountability could yield insights into effective governance counterbalances.

Methodological Expansion

Integrate additional sentiment features (e.g., tone, intensity) or data sources (e.g., earnings calls, shareholder letters) to enhance overconfidence detection models. The use of real-time transcripts from quarterly earnings calls, such as those hosted by Nvidia or Salesforce, could offer new ways to validate sentiment shifts.

Predictive Utility

Assess whether overconfidence scores predict key outcomes such as M&A activity, earnings surprises, or strategic inflection points. Case studies like Microsoft's acquisition of Activision Blizzard or Meta's pivot to the metaverse could serve as testbeds for predictive validation.

By pursuing these lines of inquiry, future research can refine theoretical models, improve empirical precision, and support the development of practical tools for behavioral monitoring in corporate contexts, particularly in the fast-evolving landscape of global technology leadership.

References

- Isaacson, W. (2011). *Steve Jobs*. Simon & Schuster.
- Kleppmann, M. (2017). *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media.
- Malmendier, U., and Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance, 60(6)*, 2661-2700.
- Malmendier, U., and Tate, G. (2005). Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited. *European Financial Management, 11(5)*, 649–659.
- Malmendier, U., and Tate, G. (2008). Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction. *Journal of Financial Economics, 89*, 20-43.
- Malmendier, U., Tate, G., and Yan, J. (2011). Overconfidence and early-life experiences: The effect of managerial traits on corporate financial policies. *The Journal of Finance, 66(5)*, 1687-1733.
- Hirshleifer, D., Low, A., and Teoh, S. H. (2012). Are Overconfident CEOs Better Innovators? *The Journal of Finance, 67(4),* 1457-1498. https://www.jstor.org/stable/23261365
- Hirshleifer, D., Low, A., and Teoh, S. H. (2012). Internet Appendix for Are Overconfident CEOs Better Innovators? *The Journal of Finance, 67(4),* 1457-1498.
- Phua, K., Tham, T. M., and Wei, C. (2018). Are overconfident CEOs better leaders? Evidence from stakeholder commitments. *Journal of Financial Economics, 127*, 519-545. https://www.elsevier.com/locate/jfec
- Zhao, Q., and Ziebart, D. A. (2017). Consequences of CEO Overconfidence. *Accounting and Finance Research, 6(2)*, 94. https://doi.org/10.5430/afr.v6n2p94
- Hatoum, K., Moussu, C., and Gillet, R. (2022). CEO overconfidence: Towards a new measure. *International Review of Financial Analysis, 84*, 102367.
- Hribar, P., and Yang, H. (2016). CEO Overconfidence and Management Forecasting. *Contemporary Accounting Research, 33(1)*, 209-240.
- Petit, V., and Bollaert, H. (2011). Flying Too Close to the Sun: Hubris Among CEOs and How to Prevent it. *J Bus Ethics, 108*, 265–283
- Kaplan, S. N., Sorensen, M., and Zakolyukina, A. A. (2021). What is CEO overconfidence? Evidence from executive assessments. *Journal of Financial Economics*. https://doi.org/10.1016/j.jfineco.2021.09.023
- Davenport, T. H., and Kudyba, S. (2016). Designing and developing analytics-based data products. *MIT Sloan Management Review, Fall 2016 Issue.* Retrieved from https://sloanreview.mit.edu/article/designing-and-developing-analytics-based-data-products/
- Dehghani, Z. (2022). *Data Mesh*. O'Reilly Media, Inc.
- Hutto, C.J. and Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.

- The New York Times. *The New York Times Developer Portal.* Retrieved from https://developer.nytimes.com/

- Google. *Gemini API documentation: Gemini model.* Retrieved from https://ai.google.dev/gemini-api/docs/models/gemini

- Moore, D. A., and Dev, A. S. (2017). Overconfidence. In V. Zeigler-Hill and T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1-2). Springer International Publishing.

- Deephouse, D. L. (2000). Media Reputation as a Strategic Resource: An Integration of Mass Communication and Resource-Based Theories. *Journal of Management, 26(6),* 1091-1112.

- Hayward, M. L. A., Rindova, V. P., and Pollock, T. G. (2004). Believing one's own press: The causes and consequences of CEO celebrity. *Strategic Management Journal, 25(7)*, 637-653.

- Phua, S. K., Tham, T. M., and Wei, Y. (2018). Overconfident CEOs and the impact on firm performance. *Journal of Business Research, 92*, 89-99.

- Galasso, A., and Simcoe, T. S. (2011). CEO overconfidence and innovation. *Management Science*, 57(8), 1469-1484. https://doi.org/10.1287/mnsc.1110.1374

- Moore, D.A., and Healy, P.J. (2008). The Trouble with Overconfidence. *Psychological Review*, *115*(2), 502-517

- Humby, C. (2006). *Data is the new oil* [Conference speech]. Association of National Advertisers.

- Palmer, M. (2011). *Data is the new oil.* Teradata Blog. Retrieved from https://marketing.teradata.com/Blog/2011/Data-Is-the-New-Oil

- Pendergraft, K. (2022). The rise and fall of WeWork: A business ethics case study. PhilArchive. https://philarchive.org/archive/PENTRA-3

- Westbrook, A. D. (2020). We('re) working on corporate governance: Stakeholder vulnerability in unicorn companies. SSRN. https://doi.org/10.2139/ssrn.3703995

Tables Index

Figures Index

Section 6 – Data Collection

Section 7 – Exploratory Data Analysis (EDA)

Section 8 – Findings and Data Product Insights

ANNEX: Supplementary Materials for CEO Overconfidence Index

Annex A: Data Sources and Corpus Overview

A.1 Article Counts and Metadata

Source	Articles Used	Columns (Metadata Fields)
NYT	2,000	9
Economist	1,876	8
EBSCO	11,795	8

Each platform provided distinct types of coverage and editorial tone. The NYT articles were more event-oriented, while The Economist supplied consistent and rich sentiment language. EBSCO contributed the highest volume, with diverse business reporting.

A.2 Sample Article Fields (Excerpt)

Source	Headline	Date	Excerpt
NYT	Google's Epic Loss, Silicon Valley's Curious Normalization	2023-12-15 10:08:28	Looks like Google is on the naughty list this holiday season...
NYT	Responsible Generative AI Is Rewriting Product Development	2023-12-13 16:16:30	In this episode, we'll explore ways businesses are rethinking AI...
Economist	Will changes to Facebook change the travel industry?	2011-09-27	Big changes are on the horizon for Facebook's travel integration...
Economist	Letters	2021-01-02	The tale of Eastern Airlines Bartleby summarises a cautionary view...
EBSCO	Will changes to Facebook change the travel industry?	2011-09-27	Big changes are on the horizon for Facebook's travel integration...

EBSCO	Letters	2021-01-02	The tale of Eastern Airlines Bartleby summarises a cautionary view...

Annex B: Keyword Taxonomy for Conf(Press)

B.1 Confidence Keywords

- visionary
- confident
- optimistic
- bold
- ambitious
- leading
- assertive
- disrupting
- transformative

B.2 Caution Keywords

- conservative
- prudent
- cautious
- restrained
- stable
- reserved
- steady

B.3 Notes on Ambiguity and Language Drift

- Certain terms such as "bold" and "ambitious" shifted meaning depending on context (e.g., bold restructuring vs. bold predictions).
- The limitations of keyword dependency highlight the importance of GenAI methods to interpret implied confidence.
- For example, the phrase "proceeded despite market skepticism" often reflected confident action but lacked direct keywords.

These keyword lists served as the lexical backbone of the Conf(Press) classifier, allowing the study to establish a comparative baseline before applying context-sensitive classification using GenAI.

Annex C: GenAI Prompt Design and Model Behavior

C.1 Classification Prompt Used

The following prompt was used with Gemini 1.0 Pro to generate consistent and context-aware classifications of CEO sentiment:

"Analyze the text of this article. Provide a 'Confident' or 'Unconfident' categorization of the CEO named in the article.
Keywords that suggest confidence may include: 'confident,' 'confidence,' 'optimistic,' 'optimism,' or others you identify.
Keywords that suggest unconfidence may include: 'pessimistic,' 'conservative,' 'cautious,' 'gloomy,' or similar.
Consider the article's tone, strategic framing, and rhetorical posture.
Summarize the rationale behind your classification."

This prompt was selected after testing multiple variations, prioritizing output stability and alignment with behavioral finance theory.

C.2 Sample Prompt Responses

Sample 1 (GenAI Output – Classified as Confident):

> "The article describes the CEO of a technology firm as 'doubling down on automation and AI,' emphasizing that their strategic vision remains unchanged despite short-term losses."
>
> → Model rationale: consistent strategic tone, assertiveness under uncertainty

Sample 2 (GenAI Output – Classified as Unconfident):

> "The article highlights the CEO's cautious messaging following weak earnings, stressing the need for measured decisions and minimizing risk."
>
> → Model rationale: emphasis on risk avoidance and limited forward positioning

C.3 Model Settings and Response Controls

- Model version: Gemini 1.0 Pro
- Temperature: 0.2 (to ensure stable output)
- Token limit: 2048
- Prompt template: Applied uniformly across articles after lemmatization and standard formatting

C.4 Interpretation Notes

- The GenAI classifier allowed for evaluation even when keywords were absent.
- It proved effective at identifying rhetorical confidence (e.g., "We remain committed to our mission") versus technical tone.
- Null or ambiguous outputs (~20% of cases) were excluded from final tallies.

Annex D: Sample Code Snippets and Classification Logic

The following code excerpts reflect the logic used to process sentiment using three distinct methods: VADER, Conf(Press), and GenAI. These are illustrative and formatted to highlight core methodological decisions.

D.1 GenAI Prompt Construction (Illustrative)

This function prepares a prompt used for zero-shot classification of CEO sentiment.

```
# Prompt Preparation Function

def format_prompt(text):
    prompt = (
        "Based on the article provided, does the CEO exhibit a confident or
unconfident tone "
        "in their actions, language, or public statements? Evaluate based
on strategic posture, "
        "leadership framing, and decision-making signals.\n\n"
        f"Article:\n{text.strip()}"
    )
    return prompt
```

Note: In production, the prompt also included keyword examples and requested a summary rationale.

See Annex C.1 for the full template.

D.2 VADER Sentiment Scoring

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
vader_score = analyzer.polarity_scores(article_text)
compound = vader_score['compound']
```

D.3 Conf(Press) Classification

```
conf_keywords = {"visionary", "confident", "optimistic", "bold",
"assertive"}
caution_keywords = {"cautious", "conservative", "prudent", "stable"}

tokenized_text = article_text.lower().split()
conf_score = sum(1 for word in tokenized_text if word in conf_keywords)
caution_score = sum(1 for word in tokenized_text if word in
caution_keywords)

label = "Confident" if conf_score > caution_score else "Unconfident"
```

These routines were executed on a document-level basis, following article filtering, normalization, and lemmatization. Additional layers such as null-checking and fallback logic (e.g., when neither keyword type appears) were included in the final pipeline.

Annex E: NLP Preprocessing Pipeline

This annex outlines the specific preprocessing steps applied to the raw article data prior to sentiment classification. These operations ensured consistency across datasets and improved classification accuracy.

E.1 Preprocessing Workflow Steps

1. Text Normalization
    - Converted all characters to lowercase.
    - Removed HTML artifacts, special characters, and URLs using regular expressions.
2. Tokenization and Lemmatization
    - Used spaCy to tokenize words and reduce them to their lemmas (base forms).
    - Ensured consistency across sources with varying editorial styles.
3. Stopword Removal
    - Applied a custom stopword list including standard English stopwords and sector-specific filler words (e.g., "company," "executive").
4. CEO-Company Association
    - Articles were matched to firms using a dictionary of ticker symbols and Bloomberg-reported CEO tenure periods.
    - Ensured each sentiment record could be linked to a specific executive and firm-year.
5. Date Filtering and Alignment
    - Matched publication dates to CEO tenure periods to avoid attributing sentiment to inactive leadership periods.

E.2 Sample Python Snippet: Lemmatization with spaCy

```
import spacy

nlp = spacy.load("en_core_web_sm")

def preprocess(text):

    doc = nlp(text)

    tokens = [token.lemma_.lower() for token in doc if not token.is_stop and
token.is_alpha]

    return tokens
```

E.3 Example Edge Case Handling

- Duplicate Removal: Articles were deduplicated using normalized titles and URL hashes.
- Ambiguous CEO Names: Used fuzzy string matching (e.g., "Tim C." vs. "Tim Cook") to ensure valid matches.
- Insufficient Content: Articles under 50 words or lacking a subject-action-object structure were excluded.

Glossary of Key Terms

Term	Definition
Conf(Press)	A press-based overconfidence index developed by Hribar and Yang (2016), which classifies CEO-related articles based on the frequency of confidence-related and caution-related keywords.
VADER (Valence Aware Dictionary and sEntiment Reasoner)	A rule-based sentiment analysis tool optimized for short texts. It generates polarity scores (from –1 to +1) to measure tone in press articles.
Polarity	A scalar value representing overall sentiment. In this thesis, it refers to the VADER compound score that captures document tone.
GenAI (Generative Artificial Intelligence)	AI models capable of generating or classifying language. Gemini 1.0 Pro is used here to assess CEO tone in articles, interpreting context beyond keywords.
LLM (Large Language Model)	A deep learning model trained on vast text corpora to perform language tasks like classification and sentiment detection.
Overconfidence	A behavioral bias where individuals overestimate their own skills, accuracy, or control. In CEOs, it can lead to aggressive or risky strategic decisions.
TDM Studio	A ProQuest platform for structured text mining. Used in this thesis to retrieve business media articles for sentiment analysis.

SIC Code (Standard Industrial Classification)	A four-digit industry code used to categorize firms. Enables sector classification into "Innovative" or "Traditional."
Bloomberg Tenure Data	Data matching CEOs to firm tenures. Used to link press articles to the appropriate executive for sentiment tagging.
Keyword-Based Method	A sentiment classification approach using predefined word lists to determine article tone. Applied in the Conf(Press) model.
Lexicon	A predefined vocabulary of terms used for sentiment tagging. VADER and Conf(Press) rely on lexicons to detect confident or cautious language.
Sentiment Analysis	A technique in NLP used to assess the emotional tone of text. This thesis uses both rule-based (VADER) and AI-driven (GenAI) sentiment classifiers.
Prompt Engineering	The design of targeted instructions (prompts) to guide the behavior of generative models like Gemini. Used to elicit confidence-related judgments from articles.