University of Amsterdam
Economics & Business

Master's Thesis in Econometrics

Innovating market sentiment analysis with natural language processing

Julian Kerssens

Master's programme: Econometrics
Specialisation: Financial Econometrics
Date of final version: August 15, 2024

Student number: 12638781
Supervisor: Dr. Y. He
Second reader: Dr. M.P. Rothfelder

Abstract

With recent breakthroughs in natural language processing (NLP) technologies, this thesis explores how these advancements can enhance market sentiment analysis. We develop a novel sentiment measure using a state-of-the-art NLP model. By creating a custom script, we collect nearly half a million keyword-specific news articles over the past decade. From this data, we construct a market sentiment index and eight asset-specific sentiment indices. The market sentiment index exhibits strong similarities and cointegration with a prominent economic sentiment measure. The asset-specific indices provide additional value by improving return predictions in well-known factor models and achieving above-benchmark returns in simple trading strategies. The minimal cost and near-continuous updating capability of these sentiment measures present an attractive alternative to traditional survey-based methods, which are often costly and slow. Furthermore, the ability to rapidly identify valuable sentiment trends offers investors, businesses, regulators, and policymakers a significant informational advantage, potentially enhancing their decision-making processes.

Statement of Originality

This document is written by Student Julian Kerssens who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

Acknowledgements

I would like to express my gratitude to my thesis supervisor, Yi He, for his mentorship, guidance and expertise throughout my research endeavor. His teachings, and course Machine Learning in Econometrics were fundamental in shaping my understanding and passion for machine learning, which eventually led me to write this thesis.

I would also like to thank my brother, Lucien Kerssens, for his insightful conversations and inspiring ideas. His enthusiasm and creative outlook have encouraged me to explore new avenues and think outside the box.

To both my supervisor and brother, I am grateful for your contributions, support and encouragement, which have significantly enriched my academic journey.

Table of Contents

1 Introduction **1**

2 Literature Review **4**

2.1 The Power of Sentiment 4

2.1.1 Shaping Economic Decision-Making 4

2.1.2 Driving Forces in Financial Markets 7

2.1.3 Early Works in Sentiment Analysis and Their Limitations 8

2.2 Contributions of Natural Language Processing 10

2.2.1 How Machines Process Language 10

2.2.2 Bidirectional Encoder Representations from Transformers 17

2.3 Applications in Market Sentiment Analysis 18

3 Methodology **21**

3.1 Data 21

3.1.1 Choice of Data Type and Wire Service 21

3.1.2 Automated Script for Data Collection 22

3.1.3 Data Processing 23

3.1.4 Data Analysis 25

3.2 The Natural Language Processing Model 32

3.2.1 Model Selection 32

3.2.2 Implementation 32

3.2.3 Sentiment Index Construction 33

3.3 Evaluating the Sentiment Indices 34

3.3.1 Cointegration and Correlation Analyses 34

3.3.2 Vector Error Correction Model 35

3.3.3 Fama-French Five-Factor Model Regressions 36

3.3.4 Sentiment Based Trading 37

4 Results **39**

4.1 The Level Indices 40

4.2 The Change Indices 45

4.3 Regression Results 49

4.4 Trading 50

5 Conclusion **54**

5.1 Giving Context to the Results 55

5.2 Contributions and Implications 56

6 Limitations and Further Research **58**

1 Introduction

Measuring sentiment and quantifying its effects is of crucial importance in economic and financial domains. For example, changes in consumer sentiment were shown to precede the onset of recessions, and predict trends in GDP (Curtin, 2004), even after controlling for the economic variables which are used to predict GDP (Howrey, 2001). In financial markets, investor sentiment was shown to affect the cost of capital of businesses (Baker & Wurgler, 2007), and be useful in constructing above-average profitable trading strategies (Tetlock, 2007).

Measuring sentiment however, is not straightforward. Traditional surveys like the Michigan Consumer Sentiment Index (MCSI) are costly to conduct, and only released monthly, usually with a delay of two weeks. These inherent limitations reduce its value, especially during economic turning-points (Shapiro, Sudhof, & Wilson, 2022). Studies have alternatively attempted to find and create imperfect measures using variables relating to trading behaviour (Baker & Wurgler, 2007), or often using textual data (Schumaker & Chen, 2009; Tetlock, 2007). While early text-based measures proved useful to some extent, language is an extremely nuanced information encoding scheme (Chen, Kelly, & Xiu, 2022), implying that intricate and sophisticated models are required to faithfully unearth information contained in text.

Improvements came as language modelling technologies advanced. Natural language processing (NLP) is a field of artificial intelligence that aims to enable computers to understand, interpret, and generate human language. This field has seen ingenious innovations in the past decade. Mikolov, Chen, Corrado, and Dean (2013) introduced methods to represent the meaning of words as vectors in high-dimensional spaces, enabling machines to unearth extremely nuanced syntactic and semantic relationships between words. A breakthrough paper with more than 125,000 citations, introduced a mechanism which empowered the integration of context into these vector representations, enabling machines to learn incredibly rich contextualized meaning of words (Vaswani et al., 2017). The integration of these mechanisms into NLP models (Devlin, Chang, Lee, & Toutanova, 2018) has given machines exceptional capabilities in understanding language, improving the proficiency in many language tasks including classifying the sentiment of lengthy texts (Chen et al., 2022; Hiew et al., 2019).

The significance and limitations in measuring sentiment, together with the recent advancements in NLP technologies, lead to the research question of this study: How can we use NLP to enhance market sentiment analysis?

This research develops an innovative method to create novel measures of sentiment, using a state-of-the-art NLP model. Through the creation of a custom-written script, I collect nearly 500,000 keyword-specific news articles from over 13,000 different media sources over a time span of more than 10 years. By aggregating the sentiment of these news articles, classified by the NLP model, we are able to create not only a market sentiment index (MSI), but also 8 asset-specific indices-an area that has been explored only sparingly in the existing literature. Figure 1 demonstrates the striking similarilities between the MCSI and MSI, which are further validated by their significant cointegration relationship.

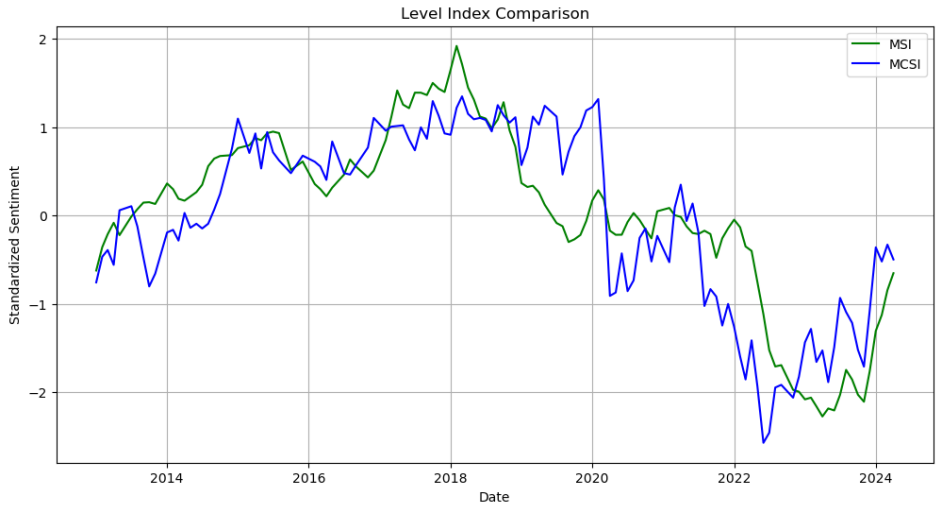

Figure 1: Comparison of prominent sentiment survey (blue) vs. NLP based sentiment measure (green).

Additional investigation, including the employment of a vector error correction model, leads to the conclusion that NLP empowers us to effectively capture the sentiment contained in news, a fundamental component which shapes consumer sentiment (Hsu, 2024).

Through a multitude of channels, it is shown how the asset-specific sentiment indices capture unique trends and offer further value over market-wide measures. Specifically, they significantly improve the return predictability in Fama-French Five-Factor model regressions, and can be used to construct above-benchmark profitable trading strategies. Figure 2 demonstrates how a simple trading strategy based solely on the sentiment indices shows improvements in multiple trading metrics against a benchmark strategy.

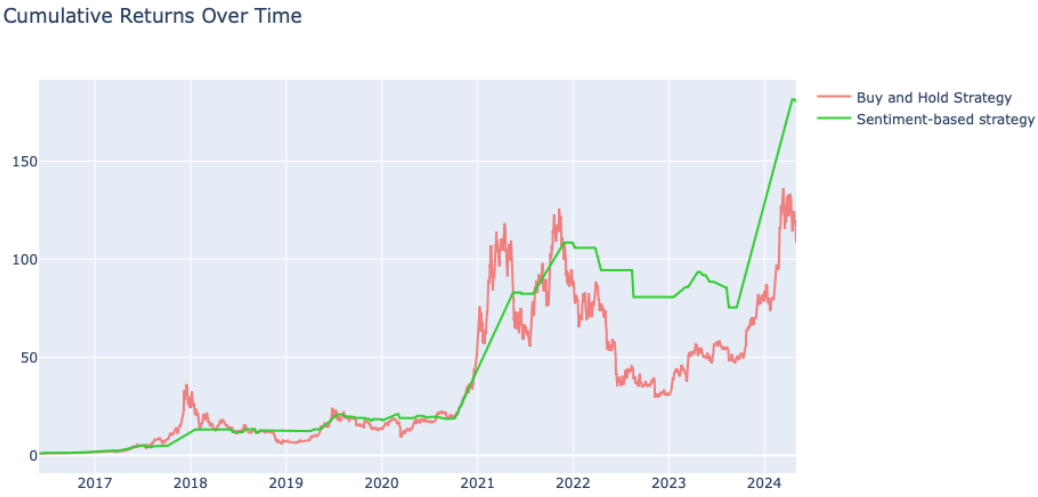

Figure 2: Comparison of cumulative returns. The sentiment-based strategy shows improvements in both maximum drawdowns and risk-adjusted returns compared to the benchmark strategy.

The findings of this research demonstrate how NLP can be used to efficiently construct low-cost, timely sentiment indices which offer attractive alternatives over traditional measures, and lead to improvements in financial forecasting applications. The ability to immediately identify valuable sentiment trends, gives investors, businesses, regulators and policymakers an informational advantage, which they can potentially use to improve their decision-making processes.

The structure of this thesis is as follows. Section 2 reviews the relevant literature, covering sentiment in the economy and financial markets, the underlying technology and contributions of NLP, and its current applications in the field. Section 3 discusses the data, detailing the selection and implementation process of the NLP model, its use in constructing sentiment indices, and the validation and application of these indices. Section 4 presents the results, discussing observations regarding the distinctions and similarities between the different sentiment indices and how they improved financial forecasting applications. Section 5 summarizes the contributions, integrates the literature and results to synthesize conclusions and answer the research question. Finally, section 6 discusses limitations, and suggests directions for further research.

2 Literature Review

In this section, we explore the existing literature and knowledge on the effects of sentiment in the economy and financial markets, with particular attention to how advancements in natural language processing (NLP) have improved market sentiment analysis. We begin by outlining the key motivators for measuring sentiment, while also addressing the challenges involved. Next, we delve into the technology behind modern NLP models, examining how they excel at understanding language. Finally, we review the current applications of these models in the literature, highlighting their effectiveness in tasks like sentiment classification. Throughout this literature review, we identify gaps that underscore the unique contributions of this research to the academic field.

2.1 The Power of Sentiment

In the context of this research, the term 'sentiment' generally refers to the overall attitude, mood, or outlook of consumers, investors, businesses, or the general public towards the economy and financial markets. Academic literature gives strong motivations for measuring and quantifying its effects. For example, Howrey (2001) found that consumer sentiment serves as a significant predictor of future GDP trends, even when accounting for the usual economic variables used in GDP forecasting. Similarly, Curtin (2004) demonstrated that shifts in consumer sentiment often anticipate changes in GDP growth and the beginning of recessions. In financial markets, Baker and Wurgler (2007) concluded how sentiment affects the cost of capital of businesses, therefore posing consequences for the allocation of corporate investment capital between firms. However, sentiment is inherently complicated to measure (Baker & Wurgler, 2007). Furthermore, many traditional measures, like surveys, are costly, slow and can suffer from biases due to small samples, limiting its practical use, especially during times of economic turning points (Shapiro et al., 2022). Diving deeper into the field, this subsection thoroughly examines these studies, and discusses how early work in sentiment analysis used NLP methods, often leading to limitations as they could not understand language effectively. These limitations underscore the need for further advancements in NLP, which are discussed in the subsequent subsection.

2.1.1 Shaping Economic Decision-Making

In the 1940s, the Economic Behavior Program was launched to explore how consumer attitudes and expectations influenced economic activity (Curtin, 2004). As part of this initiative, the Survey of Consumers was initiated in 1946 by the Survey Research Center at the University of Michigan. The survey's original goal was to gather data on household assets and debts, with little initial interest from the Federal Reserve Board (FSB) in consumer attitudes. However, the survey's founder convinced the FSB that starting with general questions about economic opinions would build trust and improve respondents' willingness to answer detailed financial questions. These general opinions turned out to be highly relevant for economic forecasting, as post-World War II income growth gave consumers more discretion in spending, often based on expectations about future income, employment, prices,

and interest rates. This relevance of consumer expectations eventually led to the creation of the Index of Consumer Sentiment (ICS), which is now known as the Michigan Consumer Sentiment Index (MCSI). This index is generally considered to be the oldest and most prominent measure for economic sentiment (van Binsbergen, Bryzgalova, Mukhopadhyay, & Sharma, 2024).

At the time of writing, Curtin (2004) had already mentioned that consumer sentiment was among the most closely monitored indicators of future economic trends, with the latest sentiment data frequently reported in the media and integrated into various macroeconomic models, including the Index of Leading Economic Indicators developed by the U.S. Department of Commerce. Figure 3 from the paper illustrates how shifts in the MCSI often preceded changes in GDP growth and the onset of recessions.

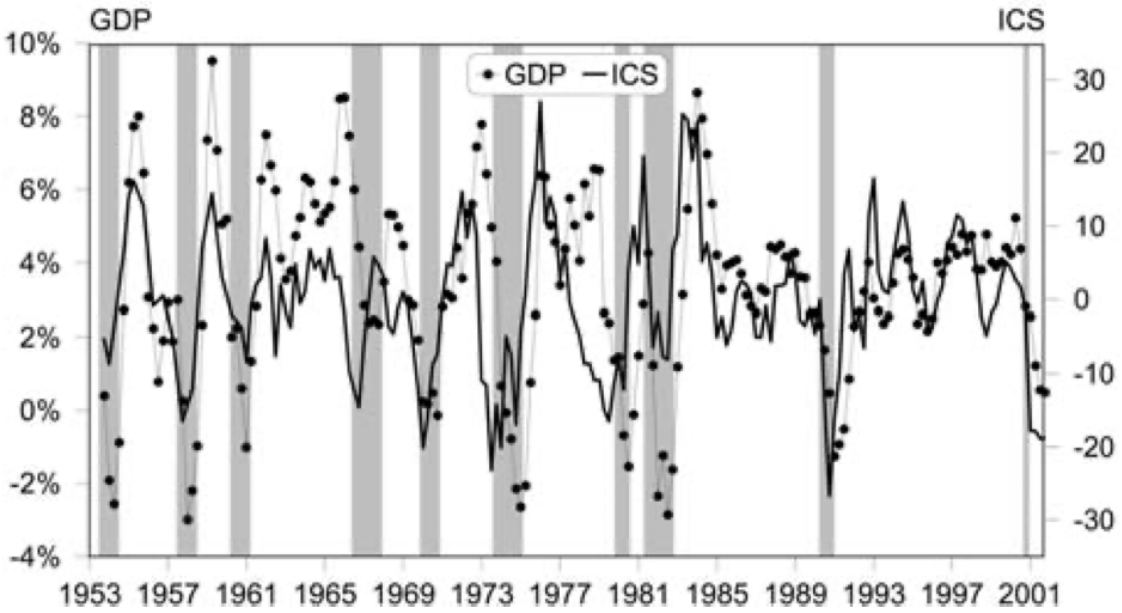

Figure 3: Consumer sentiment (ICS) and GDP time series during period 1953–2001, where the shaded areas denote periods of economic recessions.

Similarly, Howrey (2001) found that the MCSI served as a significant predictor of future GDP trends, even when accounting for the economic variables typically used in GDP forecasting.

From these observations it is evident that early insights into sentiment trends can be useful for regulators and policymakers to stabilize economic conditions and influence future economic developments. However, a notable issue with the MCSI, as with many other surveys, is its publication frequency—usually monthly—with a delay of two or more weeks (Shapiro et al., 2022). This lag can significantly reduce the usefulness of the index for policymakers and regulators, as it hinders timely understanding of sentiment trends.

To investigate whether any improved sentiment measures can be created, we must first understand the main components which form consumer sentiment. In a recent technical report of Hsu (2024), the director of the Surveys of Consumers investigates how different information sources shape

the sentiment of consumers. In figure 4, we see how news is the biggest component.

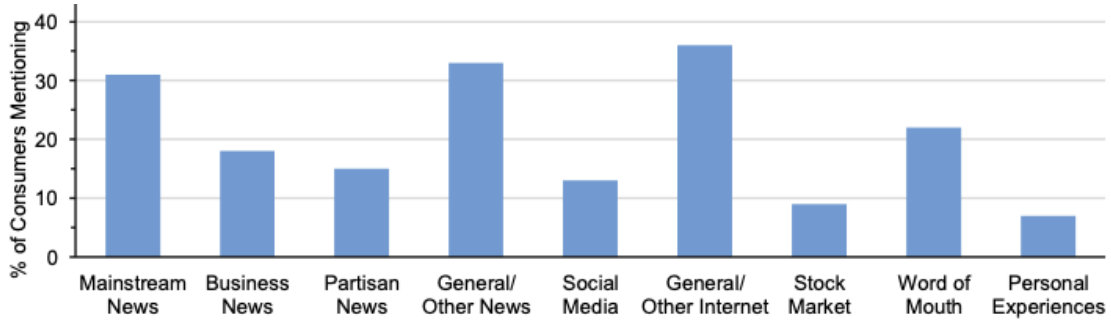

Figure 4: Shares of consumers mentioning particular information source in forming economic views.

Another figure released by the University of Michigan in 2023, further demonstrates the close relationship of news and the MCSI.

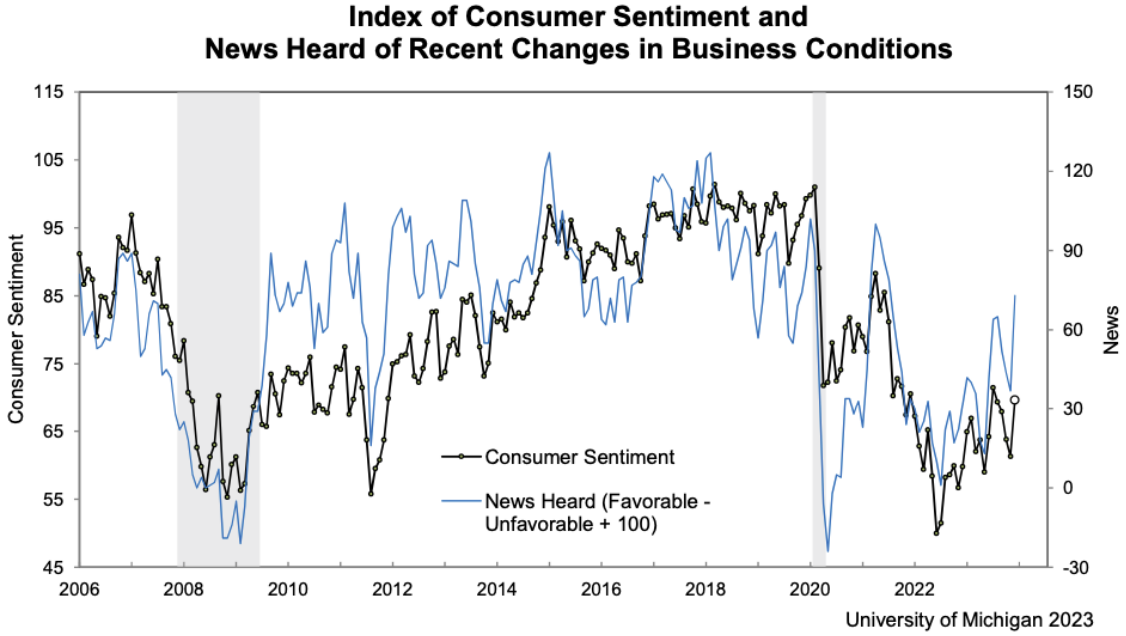

Figure 5: Consumer sentiment and news heard by consumers of recent changes in business conditions.

These observations suggest that sentiment extracted from news could be a valuable alternative for measuring consumer sentiment. Due to the abundance and continuous availability of news, it allows for capturing sentiment trends quickly and effectively.

Recent studies by Shapiro et al. (2022) and van Binsbergen et al. (2024) validate these observations, as trends in their news based measurement of market-wide sentiment lead to trends in economic fundamentals like GDP, consumption, and employment growth. Both measures also correlate with the MCSI.

This research will further investigate how news can be used to improve market sentiment analyses.

What I do differently, is creating a custom, automated script for collecting news based on specific keywords. This script allows for the collection of news articles relating to specific companies and sectors, enabling more granular analyses, not only market-wide ones. With more focus on real-time analysis, I hope to create a valuable tools for policymakers, regulators and investors alike. To understand how and through which technologies we might achieve this, we next discuss the effects of sentiment in financial markets, and how early works in the field of sentiment analysis often struggled with technological limitations.

2.1.2 Driving Forces in Financial Markets

In 1985, one of the first sentiment driven patterns was found in the stock market. De Bondt and Thaler (1985) concluded that the stock market overreacts to unexpected and dramatic news events, causing prices to adjust excessively to new information. By the early 2000's, the question was no longer whether investor sentiment affected the stock market, but instead how to measure it and quantify its effects (Baker & Wurgler, 2007).

Another major study by Tetlock (2007) found that media pessimism, extracted from the daily content of the Wall Street Journal, robustly predicted downward pressure on market prices. This pressure was followed by a reversion to the fundamental value, suggesting that the market initially overreacts to media pessimism. This finding aligned with De Bondt and Thaler (1985), and indicated that media content did not contain fundamental information about asset values, as evidenced by the significant reversion observed. Tetlock further found that these effects are exacerbated for small stocks, likely due to the fact that they have the highest individual investor ownership, and that above average profitable trading strategies could be constructed from his sentiment measure.

Similar conclusions were made by Baker and Wurgler (2007), who provided an excellent overview of popular sentiment measures of that time. By aggregating multiple proxies for investor sentiment, they constructed an index of sentiment levels, and changes. The level index captures long term trends in sentiment, whereas the changes index captures short term fluctuations. By regressing the sentiment changes index on asset returns, they are able to make some valuable additional conclusions about sentiment: Smaller, riskier assets which are harder to value, tend to be more subject to sentiment - a key observation here is that these are usually also the stocks which have a high degree of idiosyncratic variation in their returns, and that inexperienced retail or individual investors are more likely to be subject to sentiment than the professional ones. Another conclusion was that sentiment influences how risky investors perceive a company or project to be, directly impacting the expected return required by investors, and thus the cost of capital.

Works like these collectively highlighted that, besides the financial gains for investors, understanding the effects of sentiment, and being able to measure it in financial markets can be important for businesses as changes in cost of capital may force companies to delay or forgo new projects.

However, as we will see in the next subsection, early works in the field of sentiment analysis still lacked the technological advancements that we posses today, to accurately measure investor sentiment and quantify its effects at a granular level.

2.1.3 Early Works in Sentiment Analysis and Their Limitations

The term sentiment analysis gained traction in the early 2000's, and is commonly understood as the process of using natural language processing (NLP), text analysis, and computational linguistics to identify, extract, quantify, and study affective states and subjective information from textual data. Prior to this period, sentiment analysis was constrained by the lack of computational power needed to efficiently process and analyze large volumes of text data. Additionally, the availability of digital text data, such as online reviews, social media posts, and digital news articles, was limited before the rise of the internet and digital communication technologies.

Although not the initial focus, sentiment analysis found significant applications in economic and financial domains due to the previously mentioned limitations of measuring it. The growing volume of financial reports, press releases, and news articles has enabled NLP to significantly improve the modeling of various financial market dynamics, such as the effects of sentiment. This wealth of textual data, has led many businesses to leverage NLP for automated analyses, helping them maintain a competitive advantage Xing, Cambria, and Welsch (2018).

The early limitations of NLP models were highlighted in the research by Tetlock (2007) and Schumaker and Chen (2009). These studies utilized traditional feature-based methods like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Such approaches primarily relied on the frequency of individual words or simple word vectors with limited context, often failing to capture the nuanced information encoded in language.

For instance, the AZFinText system of Schumaker and Chen (2009) applied BoW and TF-IDF models to extract features from financial news articles in order to predict stock prices. While these methods were somewhat effective, they faced significant challenges, particularly with polysemy (words with multiple meanings) and a lack of contextual awareness. These limitations often led to suboptimal sentiment classification and stock return predictions.

Similarly, the research of Tetlock (2007) studied the impact of media content on investor sentiment by utilizing word counts and sentiment dictionaries to classify text into positive or negative sentiment categories. Despite providing valuable insights, this method also struggled to capture the full context and subtleties of language. For example, a phrase like "Oh great, another recession" is likely negative despite the presence of the word "great". Traditional word-based methods might misinterpret such sentences due to their simplistic nature. As noted by Chen et al. (2022), these methods are more prone to errors, especially in the presence of negation words.

Figure 6 demonstrates how a BoW vector representation is formed. Relying on solely the frequency of words in a text means that the meaning of words are considered only by themselves, completely missing how context and word order can alter the meaning of a word and the text.

Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	[1,1,1,1,2]
D2	1	1	1	1	2	[1,1,1,1,2]

Figure 6: Bag of words demonstration for two texts with identical word counts but different word orders. The method captures word frequency but loses context and word order, a limitation that can affect the accurate understanding of meaning in text analysis.

The next subsection will demonstrate how modern NLP models addressed this limitation through their capabilities to derive contextual meanings of words, meaning that they can for example recognize that the word "great", might have more of a sarcastic tone because the word "oh" is in front of it.

To conclude, we have seen how the motivations for investors, businesses, regulators, and policymakers to measure and understand sentiment have been outlined in this subsection. However, early efforts to measure sentiment in the literature were often limited by the NLP methods of that time, which struggled to accurately interpret language, thereby reducing the effectiveness of sentiment extraction from text. This highlights the necessity for more sophisticated NLP techniques in sentiment analysis.

Building on the objective of this research—to enhance market sentiment analysis through the use of news data—this study aims to further contribute to the existing literature by applying one of the most advanced NLP models available for sentiment classification of the collected news articles. By doing so, we not only demonstrate the efficacy of this underutilized model within the literature but also develop real-time sentiment measures capable of capturing unique sentiment trends related to specific assets, sectors, and the broader market.

The following subsection will discuss this NLP model and the technology underlying it, providing a comprehensive understanding of how these advanced tools interpret language and accurately classify the sentiment of extensive texts, such as news articles.

2.2 Contributions of Natural Language Processing

To fully grasp how NLP has contributed to sentiment analysis, it is essential to understand the underlying technologies that enable machines to interpret language. As previously discussed, in the early years of NLP, methods struggled to effectively capture the meaning of words due to limitations in representing linguistic context. This challenge was addressed with the development of word embeddings—a pivotal innovation in modern NLP models—that learned rich semantic vector representations of words by embedding them in a high-dimensional space.

The advent of transformer architectures marked another revolutionary leap in the field. These architectures allowed word embeddings, which were traditionally static, to be dynamically updated based on the context in which a word appeared. Bidirectional Encoder Representations from Transformers (BERT) models, built on this architecture and are particularly powerful as they can be fine-tuned for various language tasks, including sentiment classification.

Finally, we examine studies that demonstrate how these advanced technologies are being increasingly applied in the field of sentiment analysis, while also identifying areas where challenges and gaps persist.

2.2.1 How Machines Process Language

Much like a newborn baby learning a language from scratch, a machine's first essential step in language acquisition is understanding the meaning of words. The model architectures introduced by Mikolov, Chen, et al. (2013) were a breakthrough in a machine's ability to learn these meanings: The Continuous Bag of Words (CBOW) and the Continuous Skip-gram (Skip-gram) models, were both designed to efficiently embed the meanings of words into high-dimensional vector spaces, addressing earlier models' limitations regarding computational complexity.

The CBOW model predicts a target word based on the surrounding context words within a given window. It averages the vectors of the context words to predict the embedding of the target word, effectively capturing general word meanings by placing semantically similar words close to each other in the vector space. CBOW is generally faster and works well with large datasets, making it suitable for tasks where quick, general understanding of word meaning is required.

In contrast, the Skip-gram model works in the reverse direction, predicting the surrounding context words given a target word. This approach allows the model to learn more nuanced relationships between words, especially for less frequent terms. Although Skip-gram is slower than CBOW, it is more powerful for capturing a wide range of word meanings, particularly in smaller datasets, where detailed understanding of word relationships is crucial.

Given that modern large language models, such as GPT-3, use spaces with as many as twelve thousand dimensions to embed the meanings of words, understanding precisely what the model learns to encode can be challenging. This complexity arises because each dimension in the vector space captures an abstract feature of the words, which may not be easily interpretable by humans.

The Vector Offset Method (VOM), discussed by Mikolov, Yih, and Zweig (2013), is utilized to acquire insights into the syntactic and semantic relationships learned during the embedding

process. The VOM uses the concept of cosine distance as measure of similarity between the vector representations of words.

The cosine similarity between two vectors $\vec{A}$ and $\vec{B}$ is given by:

$$\mathrm{CS}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

Here, $\vec{A} \cdot \vec{B}$ represents the dot product of the two vectors, and $\|\vec{A}\|$ and $\|\vec{B}\|$ denote the magnitudes (or norms) of the vectors. Figure 7 demonstrates the concept, where a cosine similarity value close to 1 indicates that the vectors are nearly perfectly aligned, indicating that the two words have very similar meanings. A value of 0 indicates orthogonality (i.e., no similarity), and a value of -1 indicates that the vectors are diametrically opposed.

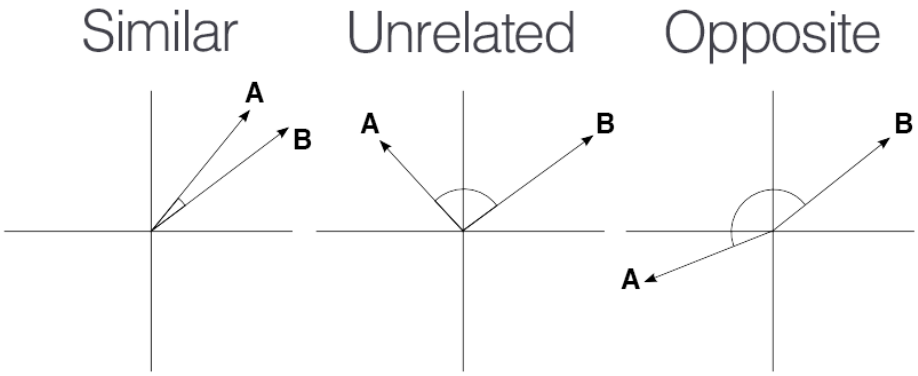

Figure 7: Illustration of cosine similarity between vectors.

Using this method, they find syntactic and semantic relations between word embeddings. For example if $\vec{E_w}$ is the vector representation of the embedding for word w, then we can search for the vector with the greatest cosine similarity to the vector obtained from the equation

$$\vec{E}_{\text{king}} - \vec{E}_{\text{man}} + \vec{E}_{\text{woman}} \quad \approx \quad ?$$

To our pleasant surprise, the answer is $\vec{E}_{\text{queen}}$. The vector space captures the concept of gender as one of the relationships embedded in its structure.

Many intriguing examples exist, for example, if we focus on the singular/plural relation, from the vector offsets we observe that

$$CS(\vec{E}_{\text{apple}} - \vec{E}_{\text{apples}}) \approx CS(\vec{E}_{\text{car}} - \vec{E}_{\text{cars}}),$$

$$CS(\vec{E}_{\text{family}} - \vec{E}_{\text{families}}) \approx CS(\vec{E}_{\text{car}} - \vec{E}_{\text{cars}}).$$

While

$$\text{CS}(\vec{E}_{\text{walk}} - \vec{E}_{\text{walking}} + \vec{E}_{\text{run}}, \vec{E}_{\text{running}}) \approx 1$$

demonstrates how verb tenses are embedded.

Figure 8 provides a simplified yet illustrative example of how word embeddings can be represented in a 3D vector space. In this visualization, different flying objects, such as "jet", "eagle", "bee", "helicopter", "drone", "rocket", and "goose", are embedded based on three dimensions: "wings", "sky", and "engine". Each object is positioned within this space according to its attributes, with coordinates indicating its specific values along these dimensions.

This example highlights the kinds of relationships and characteristics that word embeddings can capture. For instance, objects with similar functions or attributes—like "eagle", "goose", and "bee", which all have wings and are commonly associated with the sky—are placed closer together in the vector space, reflecting their shared features. On the other hand, mechanical flying objects like "jet", "helicopter", "drone", and "rocket" are distinguished by their reliance on engines, which shifts their position along the "engine" dimension.

This visualization also hints at the unexpected and sometimes non-obvious relationships that embeddings might learn. For example, "jet" is represented as having moderate values across all three dimensions and it exists in a position that blends characteristics of both natural and mechanical flying entities. This demonstrates how embeddings might capture nuanced or hybrid attributes that arise from contextual similarities during training rather than from strict categorical definitions.

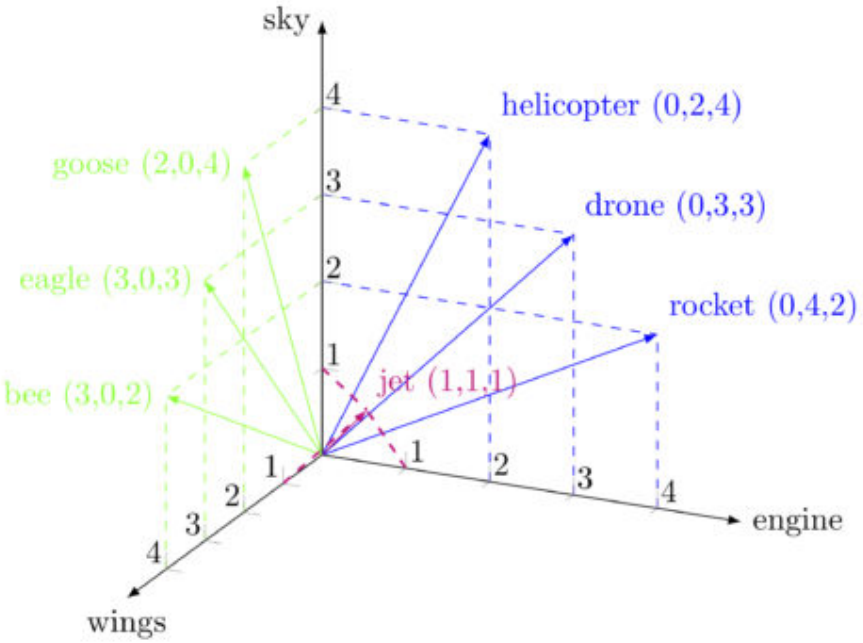

Figure 8: Example 3D vector space visualization with word embeddings for flying objects. Each dimension represents an abstract feature learned from contextual data. The visualization illustrates how different attributes, such as 'wings', 'altitude', and 'engine dependency', influence the vector representation of words like 'jet', 'eagle', and 'drone' in the embedding space.

In this 3D example, the simplicity of the dimensions allows us to easily visualize and understand how these relationships might work in higher dimensions. However, when the number of dimensions increases—such as in the hundreds or thousands used by modern word embedding techniques—the relationships captured by the word embeddings become far more complex and less intuitive. Each additional dimension can represent increasingly abstract features of the words, which might not correspond directly to any easily identifiable characteristic. For instance, in a high-dimensional embedding space, one dimension might capture a subtle aspect of a word's connotation, such as the degree of formality. As a result, words like "house" and "home" might be positioned close to each other due to their similar meanings, but a specific dimension might slightly shift "home" away from "house" due to its warmer, more personal connotation. This shift may not correspond to any clear, singular attribute like "location" or "size", but instead reflects a more abstract quality that is learned from the context in which these words typically appear.

Although word embeddings were revolutionary in enabling machines to capture the complex and nuanced meaning of words, they still had significant limitations, with one of the primary shortcomings being their inability to integrate context. Consider the two sentences: "This stock is trading at a premium", versus "Insurance premiums are rising". In these examples, the word "premium" clearly conveys different meanings depending on the context, yet traditional word embeddings would assign the same vector representation to "premium" in both sentences.

Limitations like these posed hindrances in language tasks, such as sentiment classification. In the first sentence, "premium" carries a positive sentiment, as it suggests that the stock is highly valued. Conversely, in the second sentence, "premium" has a negative connotation, as rising insurance premiums imply higher costs for consumers. Since traditional word embeddings assign the same vector to "premium" in both contexts, the model might incorrectly interpret the sentiment of these sentences. For instance, it might mistakenly perceive "Insurance premiums are rising" as a positive statement, simply because the word "premium" is generally associated with positive meanings in other contexts.

Furthermore, word embeddings struggled with understanding phrases or multi-word expressions that have specific, non-compositional meanings, often referred to as idiomatic phrases (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Consider the sentence, "The company's CEO said the new policy will help us weather the storm". Here, the phrase "weather the storm" is an idiom meaning to endure a difficult situation or crisis. A sentiment classification model using static word embeddings can't recognize the idiomatic nature of this phrase and might interpret "storm" literally, associating it with negative connotations like chaos or disaster. As a result, the model might incorrectly classify the sentiment of the sentence as negative, failing to capture the positive or resilient sentiment intended by the speaker. This limitation arises because the vector representations of words typically do not consider word order or the interaction between words within a phrase, leading to a loss of information which can be crucial for understanding the full meaning or sentiment of such expressions.

Continuing with our analogy, as a child grows, they start to develop a deeper understanding of how words relate to each other within sentences. For example, a child will learn that the meaning of the word "mouse" changes depending on whether it's used in "The mouse ran across the floor" or "I need a new mouse for my computer". This step in language acquisition involves understanding context and the dynamic nature of meaning.

In a similar way, the Attention Mechanism (AM), originally popularized by Bahdanau, Cho, and Bengio (2014), represented the first significant step in enabling machines to understand and leverage context, specifically for the task of language translation. This mechanism was particularly powerful in preserving the meaning and context of the translation, ensuring that even as the sentence structure changes between languages, the core message remains accurate, enabling more fluent and contextually aware translations.

However, the underlying architecture of AMs—Recurrent Neural Networks (RNNs)—posed significant limitations for many language tasks. First, RNNs struggled with processing long sequences of text efficiently because they process sequences step by step, making them inherently slow and computationally expensive, particularly for large texts, and had difficulty learning long-range dependencies within these texts. Second, RNNs were not well-suited for parallelization, a critical feature for handling the extensive amounts of text data often encountered in NLP, such as entire corpora of books, web pages, or social media content. Parallelization allows large datasets to be processed more quickly by distributing the workload across multiple processors, significantly speeding up both training and inference. Due to their sequential nature, RNNs could not be easily parallelized, mak-

ing the processing of extensive datasets much slower and severely limiting the potential of NLP models to scale efficiently. As a result, earlier models struggled to perform well on tasks requiring the analysis of large amounts of text data, such as language translation or sentiment classification.

Transformers were a breakthrough not only in addressing these limitations but also in enabling the updating of the meaning of words based on their context within a sentence. The seminal paper by Vaswani et al. (2017), which introduced the transformer architecture, has garnered over 125,000 citations in less than a decade. The key innovation in the architecture was the introduction of "self-attention", a mechanism that allows the model to focus on the relationships between all words in a sentence to update the vector representation of each word embedding based on its context. Moreover, the transformer architecture is highly parallelizable and efficient for processing large texts. These advancements meant that advanced language tasks, like sentiment classification, which require contextual understanding of large texts, could now be completed more accurately and effectively.

As our final goal is to understand how the underlying technology of modern sentiment classification models works, we discuss the mathematics behind the attention mechanism for updating word embeddings based on their context within a sentence.

Figure 9 provides a simplified visual example of how we might expect attention to function, showing how certain words in a sentence "attend" to others. The arrows indicate the direction of attention, illustrating which words contribute to refining the meaning of others within the sentence.

Figure 9: Simplified illustration of the self-attention mechanism. The highlighted words are identified as being relevant to update the meaning of other words in the sentence. The arrows indicate how for example the adjectives "fluffy" and "verdant" influence the understanding of the nouns "creature" and "forest", respectively.

To understand the exact mechanics of the updating mechanisms, first consider the something called the query matrix M_Q. M_Q can be thought of as using numbers (weights) to encode "questions" which it will ask each word in a sentence. For illustrative purposes, it might encode the question "are there any adjectives before this word embedding which significantly alter its meaning?" The key matrix M_K, can be thought of the optimal way to answer the questions in M_Q.

By multiplying each word embedding $\vec{E}_i$ by the query matrix M_Q, we obtain the query vector for that embedding $\vec{Q}_i$. Similarly, we obtain key vectors through $\vec{K}_i = M_K \vec{E}_i$.

By taking the dot product of each query and key vector pair QK^T, we obtain a matrix representing the similarity between the query and key vectors. The alignment of these vectors, as measured by their dot product, determines the importance of one word in updating another. Specifically, $QK^T_{i,j}$, represents the i'th word's importance in updating the j'th word. We then apply a softmax transformation to each column, which normalizes these values into probabilities that sum to 1.

These probabilities, or weights, indicate the significance of each word in updating the meaning of the others. Figure 10 illustrates what an "attention pattern" or "attention heatmap" might look like. We see, for instance, that the adjectives "fluffy" and "blue" are important for refining the meaning of the noun "creature", whereas the word "blue" for example does not significantly influence the meaning of "forest". As expected, the self-attention mechanism effectively only focusses on relevant contextual words to update the meaning of other words.

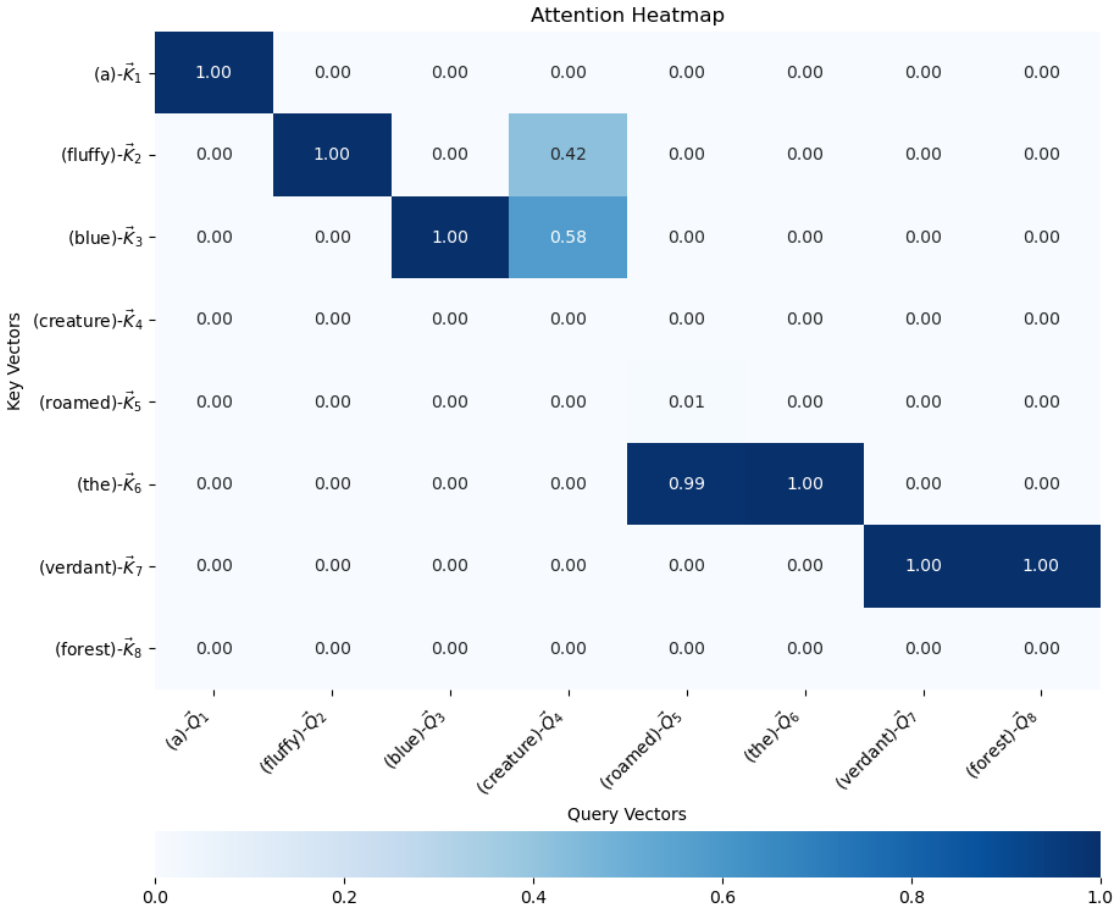

Figure 10: Attention heatmap. Each column demonstrates the weights of each word in updating the meaning of the word of that column.

Now that we know the attention pattern, we must update each word embedding. For this, a value matrix M_V is used. In a similar way to M_Q and M_K, we multiply each embedding by the value matrix to obtain value vectors: $\vec{V}_i = M_V \vec{E}_i$. Finally, the attention values in each i'th entry of column j is multiplied by $\vec{V}_i$. The sum of the j'th column of these weighted value vectors represent the update made to $\vec{E}_i$

The previously discussed mechanics can be compactly written in the form

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

where $\sqrt{d_k}$ is introduced for numerical stability, but not important for the intuitive interpretation.

Figure 11[1] intuitively demonstrates how the final Attention(Q, K, V) matrix updates the embedding of a word.

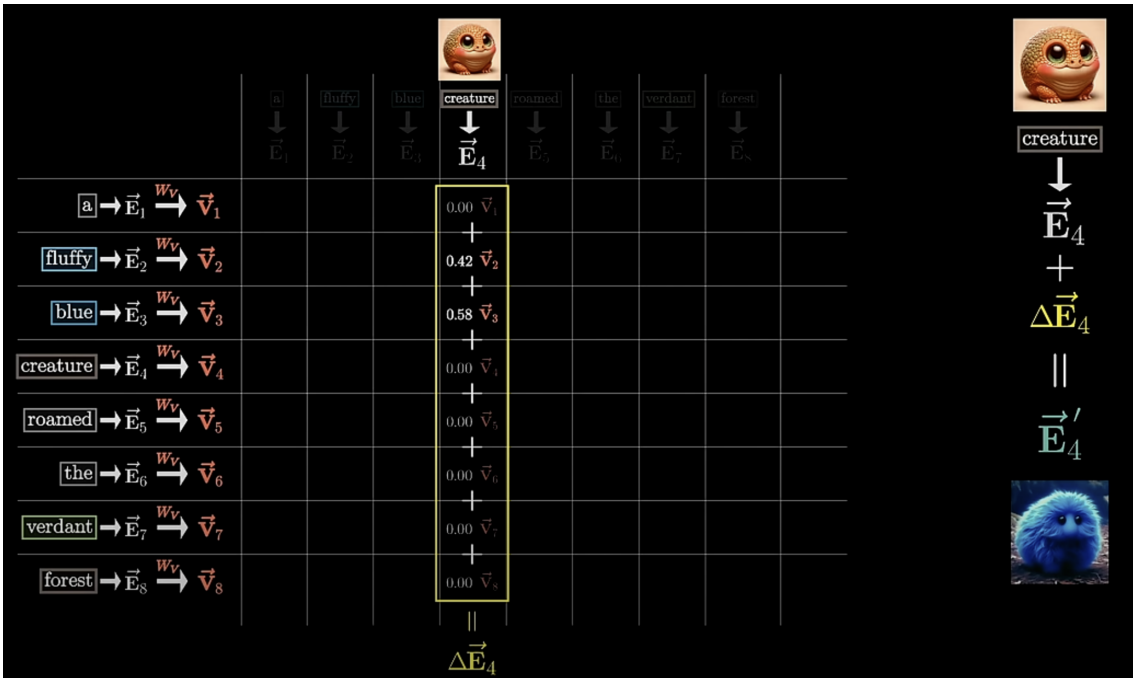

Figure 11: Example of how the word embedding for "creature" is updated within the context of the example sentence. The visual represents how attention weights, computed from other words like "fluffy" and "blue", contribute to adjusting the embedding vector $\vec{E}_4$ corresponding to "creature".

By now we have an understanding of how transformer architectures work. Next, we discuss one of the most popular modern models which integrates the transformer architecture and can be fine-tuned for very many language tasks.

2.2.2 Bidirectional Encoder Representations from Transformers

The integration of transformers into NLP models lead to significant advancements in sentiment analysis. In 2018, Google introduced the Bidirectional Encoder Representations from Transformers (BERT) model, as presented by Devlin et al. (2018). BERT was notable for its exceptional proficiency in classifying sentiment due to its ability to understand bidirectional context. It achieved accuracies as high as 94.9% on tasks like the Stanford Sentiment Treebank, part of the General Language Understanding Evaluation benchmarks (Hiew et al., 2019). Despite its effectiveness, BERT's computational demands, with models often exceeding one hundred million parameters, made it resource-intensive and impractical for all applications. In 2019, the Robustly Optimized BERT approach (RoBERTa) was developed by Liu et al. (2019). RoBERTa optimized the pre-training process of BERT and utilized more data, leading to improved performance. These optimizations allowed

[1] `AttentioninTransformers`

RoBERTa to outperform BERT in various NLP tasks. Later in 2019, Sanh, Debut, Chaumond, and Wolf (2019) introduced DistilBERT, a distilled version of BERT designed to be smaller and faster while retaining much of BERT's accuracy. DistilBERT used knowledge distillation, where a smaller student model is trained to mimic a larger teacher model. This concept of distilling BERT models made them more efficient in terms of computational resources while still delivering high performance.

As shown in Figure 12, in recent years, we have seen an explosive increase in transformative technologies for NLP tasks. As a consequence of their recent emergence, many of these technologies remain underexplored in the literature. To address this gap, this research applies a state-of-the-art distilRoBERTa model for sentiment analysis. The next subsection explores current applications of modern NLP models in sentiment analysis and financial forecasting, highlighting key discoveries and identifying areas where further conclusions are yet to be drawn.

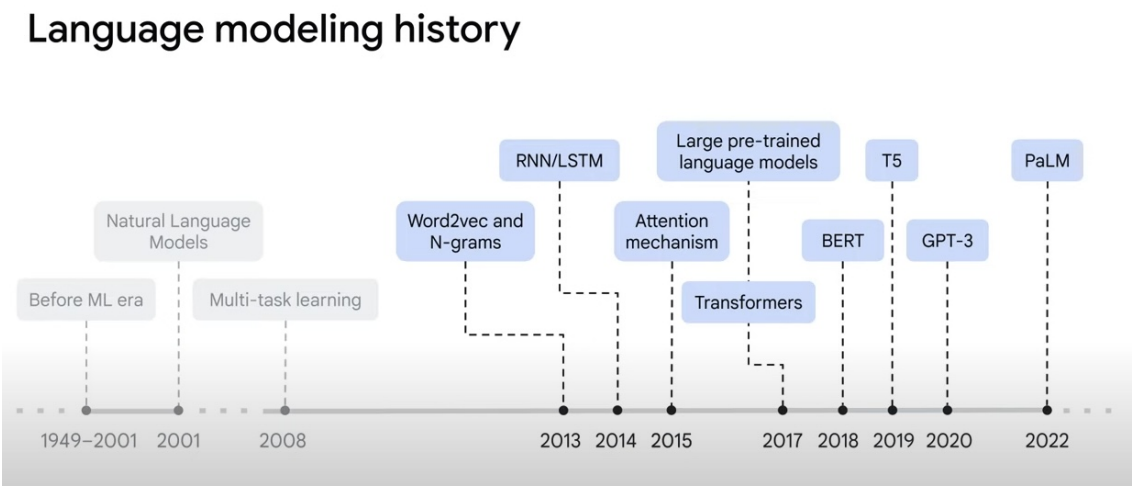

Figure 12: A timeline illustrating developments in the field of language modeling. Starting from early developments in natural language processing, the timeline showcases how advancements have accelerated dramatically in recent years.

2.3 Applications in Market Sentiment Analysis

With a solid understanding of modern NLP models, as well as the significance of sentiment in financial and economic contexts, we are now well-equipped to review the current literature. Having already outlined specific contributions in earlier subsections, we will further examine related works in a comparative manner, positioning this research within the broader field and highlighting its unique relevance to market sentiment analysis.

Initial motivations for applying advanced deep learning technologies in financial markets were given by Fischer and Krauss (2018), who obtained extreme returns and significantly outperformed traditional methods with their trading strategies. Chen et al. (2022) further investigated the predictability of stock returns induced by global news across 16 international equity markets, employing the latest large language models such as BERT and generative pre-trained transformer models. Their findings revealed that information from newswires is incorporated into asset prices with an ineffi-

cient delay. This delay can be exploited through real-time trading strategies, demonstrating that the sentiment embedded in news articles can serve as a potent predictor of future asset price trends. They further conclude that information contained within news texts is not immediately and fully integrated into market prices. This lag may be due to factors such as limits-to-arbitrage (Shleifer & Vishny, 1997), induced by noise trader risk (De Long, Shleifer, Summers, & Waldmann, 1990). As Barberis and Thaler (2003) noted, there are an increasing number of retail investors entering equity markets and these investors are often naively and insufficiently diversified, leaving them sensitive to unsystematic risks which are often covered in the news. Besides these facts, we have learned how retail investors are most subjective to erratic behaviour, making it harder for arbitrageurs to trade in markets where many retail investors are present. The study of Chen et al. (2022) additionally highlighted that news sentiment is particularly effective in predicting future returns for small stocks. Several economic explanations were provided for this observation:

- **Attention Deficit**: Small stocks receive less attention from investors, leading to slower responses to news.
- **Uncertain Fundamentals**: The underlying fundamentals of small stocks are more uncertain and less transparent, requiring more effort to process news and translate it into actionable price assessments.
- **Liquidity Constraints**: Small stocks are less liquid, necessitating more time for trading to incorporate information into prices.

These findings were in line with Baker and Wurgler (2007), highlighting that even after 15 years and using more sophisticated models, similar sentiment induced market patterns were still present in financial markets. This suggests that these observations are not temporary anomalies but persistent features in the market.

Additionally, Chen et al. (2022) discuss how research has so far explored only a limited portion of market-relevant textual data, often concentrating on specific data sources like the front page of The Wall Street Journal or the "risk factor" section of 10-K filings, indicating that text data remains an underexploited data source for understanding asset markets.

Many of the major studies, who do use a diverse set of textual data in their sentiment analyses, such as Baker and Wurgler (2007), Shapiro et al. (2022), and van Binsbergen et al. (2024) however, focus primarily on market-wide analyses. These studies lack granular insights into the foundations and variations in investor sentiment in different markets over time and do not determine which specific stocks attract speculators or have limited arbitrage potential, which was one of the main future challenges described by (Baker & Wurgler, 2007).

It was not until 2019 that we saw some of the first studies using BERT models to construct asset-specific sentiment indices (Hiew et al., 2019), who noted how the literature had underexplored the use of state-of-the-art NLP models for sentiment analysis in financial markets. Their indices demonstrated significant return predictability. However, their analysis was limited to only three

assets in Chinese markets and did not discuss the underlying economic mechanisms driving their results. This was another limitation earlier highlighted in the work of Baker and Wurgler (2007), who explained that even when significant return predictability is found, elucidating the economic reasoning remains a challenge.

In 2023, Kelly and Xiu (2023) emphasized that news text was still a underutilized data source in empirical models, mentioning that asset price variation was expected to predominantly eminate from the arrival of unanticipated news. Xing et al. (2018) additionally mentioned how research indicates that that a combination of subjective sentiment and objective event facts would take advantage of each other and product even better forecasting results. Observations like these further motivate the idea of using news text in financial forecasting applications as news articles present objective facts but also often convey subjective sentiment, either through the language used or the interpretation of events. Finally, Xing et al. (2018) highlights how 'information overload' among market participants can lead to information asymmetry, creating opportunities for excess returns. This provides strong motivation for using the earlier mentioned distilRoBERTa model in financial forecasting, as its efficiency in processing large volumes of text quickly may enable us to capitalize on this information asymmetry and potentially generate profits.

To conclude, recent literature clearly demonstrates the continued relevance of sentiment analysis, with similar sentiment driven patterns still present in markets more than a decade after first being observed, suggesting it is a persistent feature of economic and financial domains. Many studies have successfully used the latest NLP technologies to construct market sentiment measures that predict economic fundamentals and improve return predictions. However, there remains a gap in providing more granular insights into how sentiment affects specific asset markets, and some studies lack an economic understanding behind their sentiment driven results. Additionally, motivations are given for further exploring news text data and using models capable of handling large volumes of it, leading to advantages in financial forecasting due to advanced information processing capabilities.

The methodology of this research was meticulously devised to address these identified gaps in the literature. Through the previously described custom written script, I collect nearly 500,000 news articles, from more than 13000 different sources. Using a novel, state-of-the-art NLP model designed specifically for efficiency and handling large amounts of data quickly, I construct 8 asset specific sentiment indices and one market sentiment index. These indices allow for the timely capturing of unique sentiment trends, significantly improving trading strategies, and return predictability in Fama-French Five-Factor model regressions. Moreover, the market sentiment index aligns closely with long-term trends observed in the Michigan Consumer Sentiment Index (MCSI), showing a strong correlation of 82%. Through advanced econometric methods like cointegration tests and vector error correction models, I investigate the economic intuition behind the results. Not only do these these indices and analyses allow for synthesizing more granular insights about how sentiment affects different markets, they also offer a fast, cost effective alternative to traditional measures like the MCSI.

3 Methodology

This section details the systematic approach for collecting and transforming news articles into valuable sentiment data, as well as the methods by which this data is effectively interpreted and applied in financial forecasting. We begin with an in-depth discussion of the data collection process, including the selected data type, and the development of a custom script to gather thousands of articles over time. Next, we outline the preprocessing steps necessary to prepare the data for natural language processing (NLP) models, followed by a comprehensive analysis of the dataset. We then explore the selected NLP model, its implementation for sentiment classification, and the construction of sentiment indices. Finally, we describe the methods used to analyze these indices and apply them in financial forecasting applications.

3.1 Data

Before collecting the extensive data required for constructing sentiment indices through NLP, two primary decisions need to be made: first, determining the most appropriate type of sentiment data; and second, selecting a suitable method or wire service for data collection.

3.1.1 Choice of Data Type and Wire Service

With the initial motivations discussed in the previous section, news articles are chosen as the primary source of sentiment data for several reasons.

Firstly, news articles can provide in-depth analyses and detailed information on market events, company performance, and economic indicators. This information is highly relevant for forecasting market movements, making news articles a valuable source of sentiment data (Tetlock, 2007).

Secondly, news articles are available in abundance and from many sources, which is crucial for creating aggregated sentiment indices. This abundance ensures a comprehensive dataset that can capture a wide range of market sentiments and trends over time. In contrast, financial and earnings reports, although valuable, are not published daily, making them less suitable for continuous sentiment analysis (Xing et al., 2018).

Finally, while social media posts, such as those from Twitter, have been successfully integrated into models to improve predictions in financial markets (Bollen, Mao, & Zeng, 2011), and are also available in abundance, they often contain significant noise and bias due to their informal nature and the diversity of user opinions (Xing et al., 2018). News articles, on the other hand, are typically written by professional journalists and analysts, which helps maintain a higher standard of objectivity and credibility. Furthermore, they are published at frequent time intervals, which is an important factor according to Xing et al. (2018). By using news articles, this study aims to leverage more reliable and consistent sentiment indicators and reduce the impact of biases.

When it comes to data collection, there are many wire services available for gathering news articles. Unfortunately, most of these services are expensive, purposely delay the collection of live news, and only allow for a limited number of requests within given time intervals. Thankfully, there

are also open-source packages that are free to use. The GNews package on GitHub[2] is a Python library designed to interface with Google News[3]. It allows users to easily search for news articles based on specific keywords and retrieve these articles in JavaScript Object Notation (JSON) format by providing an application programming interface. This library, together with an automated script described in the next subsection, is used for the data collection process.

3.1.2 Automated Script for Data Collection

To construct sentiment indices over time, a continuous amount of articles is needed. Given that the GNews library does not directly support sending multiple requests over multiple dates, an automated script was created. When running the script, the user is prompted to specify a keyword for which news should be collected, and a time-frame over which the script will iteratively send requests. Since Google News appears to limit each request to a maximum of 100 articles, it is advisable to select a minimum time-frame of two days to obtain the most amount of data. The script then asks for a starting date, which requires an input in the format YYYY-MM-DD, or simply the word "now" to start from the current date. The script automatically stops running once it retrieves zero articles three times in a row.

For example, using the inputs "Ethereum", "2", and "now", the script requests all articles from Google News over the past 2 days that contain the word "Ethereum" in the title. Once retrieved, to prevent data loss if the script accidentally stops the articles are immediately appended to a file called "Ethereum.json". This file is created in the folder where the script is run. The script then sends the same request for the 2 days prior to the past request and then appends these articles to the same file. This process continues until no more data is retrieved. Figure 13 illustrates this process.

[2]https://github.com/ranahaani/GNews
[3]https://news.google.com/home

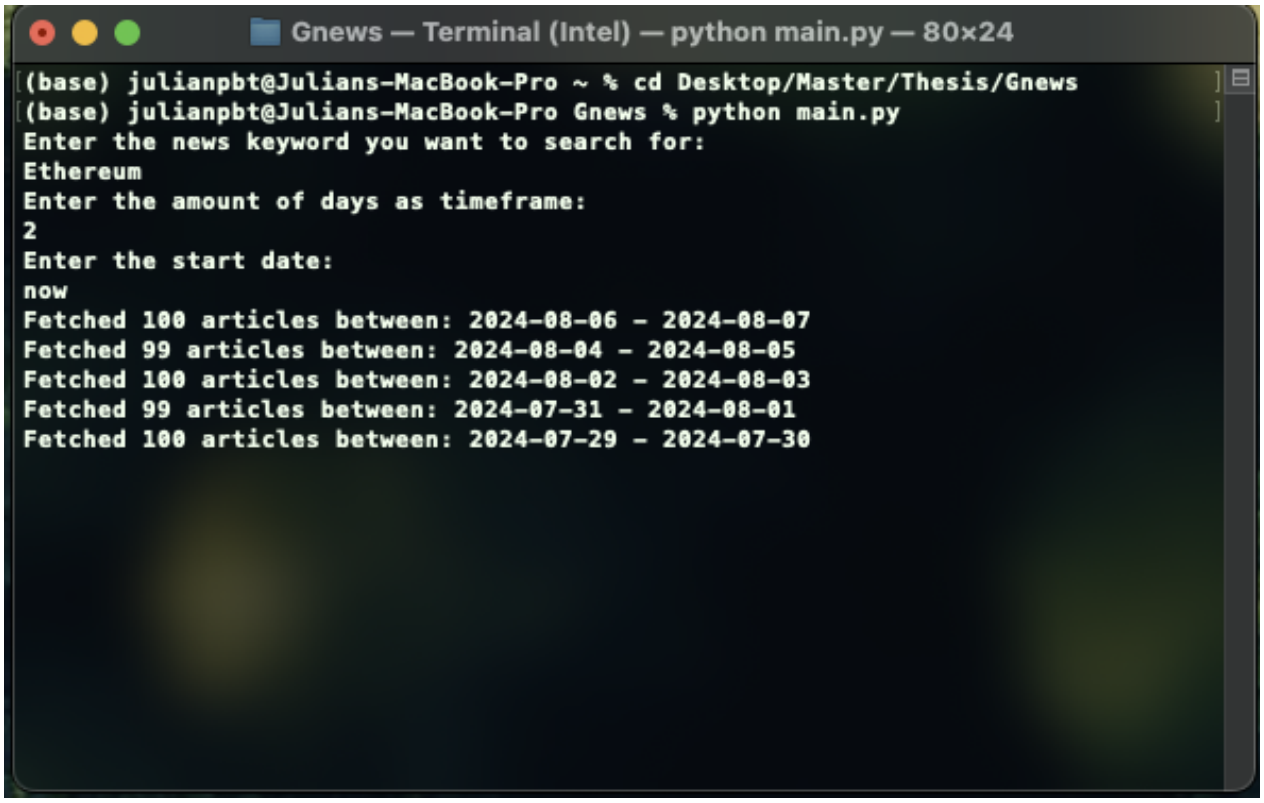

Figure 13: Illustration of the automated script for collecting data.

3.1.3 Data Processing

When collecting the data as described in the previous subsection, a choice had to be made regarding how to append articles to the JSON file. When the first set of articles is retrieved, they are added to a Python array so that they can be immediately imported into Python. Because the automated script collects multiple arrays, simply appending them together in one file will break the JSON syntax and therefore cannot be directly loaded in to Python.

The alternative is to overwrite the previously fetched articles with the new ones together with all the old ones to maintain the correct JSON syntax. However, this method requires all existing articles to be kept in memory. As the files are usually upwards of 500MB, this poses a significant risk of losing data if a crash occurs. Therefore, the first option was implemented. Figure 15 presents an example of some raw data, obtained using this method.

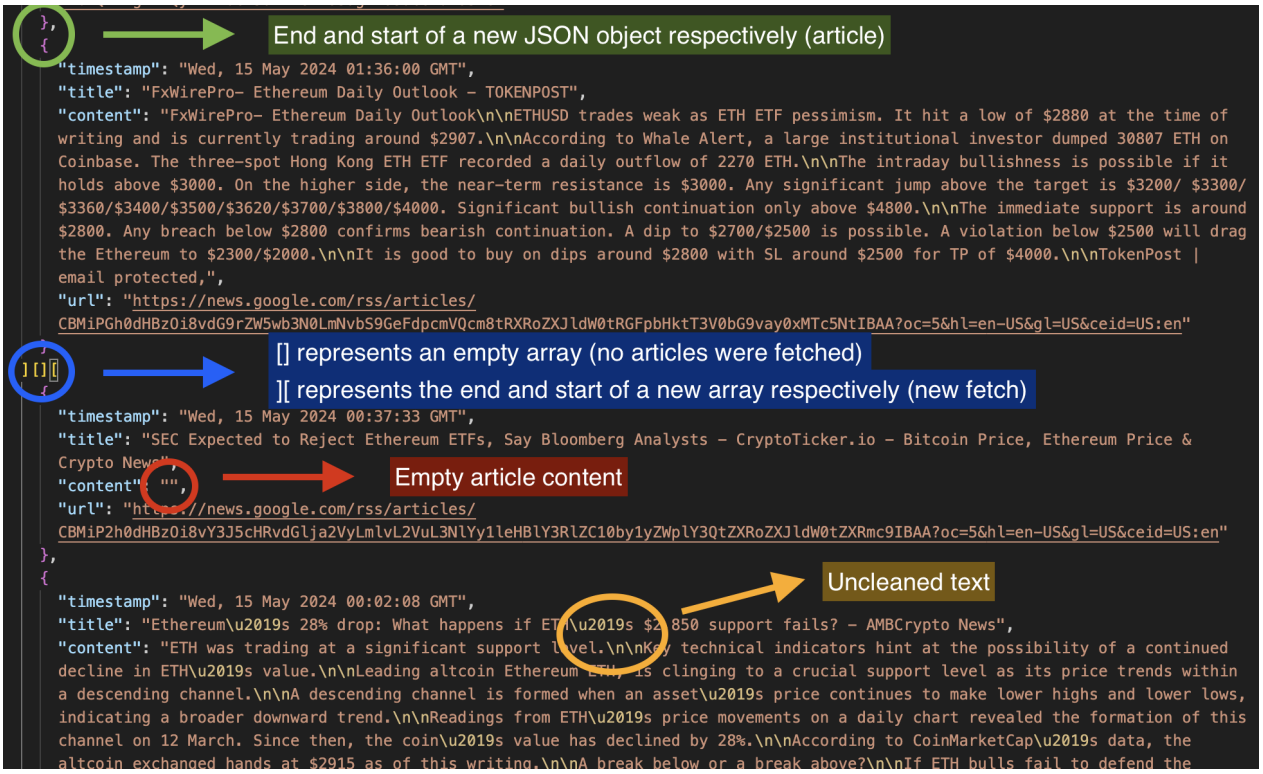

Figure 14: Unprocessed Example Data. This figure shows a raw JSON file containing unprocessed data for Ethereum-related articles. Key elements are highlighted: the green circle highlights the curly braces that indicate the beginning and end of a JSON object, representing an individual article; the blue circle highlights the square brackets [],][, which denote the start and end of new arrays, indicating new fetches of articles and a fetch where no articles were retrieved; the red circle highlights empty article content, showing instances where the article content could not be retrieved; and the orange arrows point to uncleaned text within an article, demonstrating the presence of unnecessary characters. This raw data must be cleaned and processed before it can be effectively used for sentiment analysis.

Because a new array is appended to the file each time we fetch a new set of articles, we are left with multiple arrays which is not correct JSON syntax readable by Python. We need to make sure the file is a single array of shape [{article},{article}, ... ,{article}], where {article} is a JSON object pertaining information about a single article.

To achieve the correct syntax, some additional steps must be taken. First of all, all empty arrays ("[ ]"), corresponding to a timeframe where no articles were retrieved, as seen in the blue circle of figure 15 must be removed. Furthermore, the opening and closing brackets of each intermediate array ("] [") must be replaced with a comma (",").

Finally, articles with empty content, as seen in the red circle of Figure 15, are removed. The article contents are then cleaned using Flair's `Sentence` class, which is designed to handle text for NLP tasks. This class processes and transforms raw text into a cleaner format suitable for sentiment classification. The cleaning process involves removing irrelevant characters and elements, such as HTML tags, newline characters, and other extraneous symbols that add unnecessary noise to the

model's analysis. Exact details about the cleaning and processing can be found in the documentation for this class on GitHub.[4] (line 703). As an example, figure 15 shows how the title in the orange circles of figure 14 is cleaned.

```
1 from flair.data import Sentence
2 Sentence("Ethereum\u2019s 28% drop: What happens if ETH\u2019s $2,850 support fails?\n\n",
3          use_tokenizer=SegtokTokenizer())
```

```
Sentence[16]: "Ethereum's 28% drop: What happens if ETH's $2,850 support fails?"
```

Figure 15: NLP data cleaning example.

3.1.4 Data Analysis

I collected news articles based on the following keywords: "Dow Jones", "JP Morgan", "Amazon", "Apple", "Nvidia", "Boeing", "Tesla", "Bitcoin", and "Ethereum". Because Tetlock (2007) finds that measures of news which cover primarily stocks in the Dow Jones index serve as a proxy for overall investor sentiment, news articles based on the keyword "Dow Jones" are used to construct the market sentiment index of this research. Intuitively this is justified by the fact that articles related to the Dow Jones cover 30 of the largest and most influential companies in the U.S., which can capture a broad view of market sentiment that reflects the overall health of the stock market.

Figure 16 below illustrates the yearly number of collected articles over time for the various datasets. The y-axis is plotted on a logarithmic scale to accommodate the wide range of article counts. Each line represents the number of articles per year for a different dataset based on the keyword that was searched on.

Intuitively, there are several desirable properties that we hope to observe in our dataset after collection. Firstly, having a diverse set of sources is beneficial. As noted by Chen et al. (2022), previous studies often rely on a single data source, making analyses susceptible to the biases of that particular media distributor. Additionally, it would be advantageous to have a greater number of articles from more reputable sources. Ideally, the dataset would naturally skew towards more credible sources rather than having an equal number of articles from each media source regardless of their credibility. This is more useful for constructing sentiment indices, as credible sources are more likely to provide accurate and reliable information, thereby enhancing the validity and robustness of the sentiment analysis.

Table 1 and Figures 17 and 18 collectively demonstrate how these properties have been achieved. Table 1 presents statistics of the diverse dataset containing nearly 500,000 news articles from more than 13,000 different media sources and provides initial insights into the skew of the data. The large mean but small median indicates that a few outlets have contributed a disproportionately high number of articles. Figure 17 plots this distribution of the total numbers of articles collected per media outlet, confirming this observation. Roughly 25% of all articles are from the top 15 sources in the dataset. It is crucial that these top sources are reputable and credible. Figure 18 presents a pie chart of the top 15 sources, showing that even among these sources, the distribution is relatively

[4] https://github.com/flairNLP/flair/blob/v0.13.1/flair/data.py

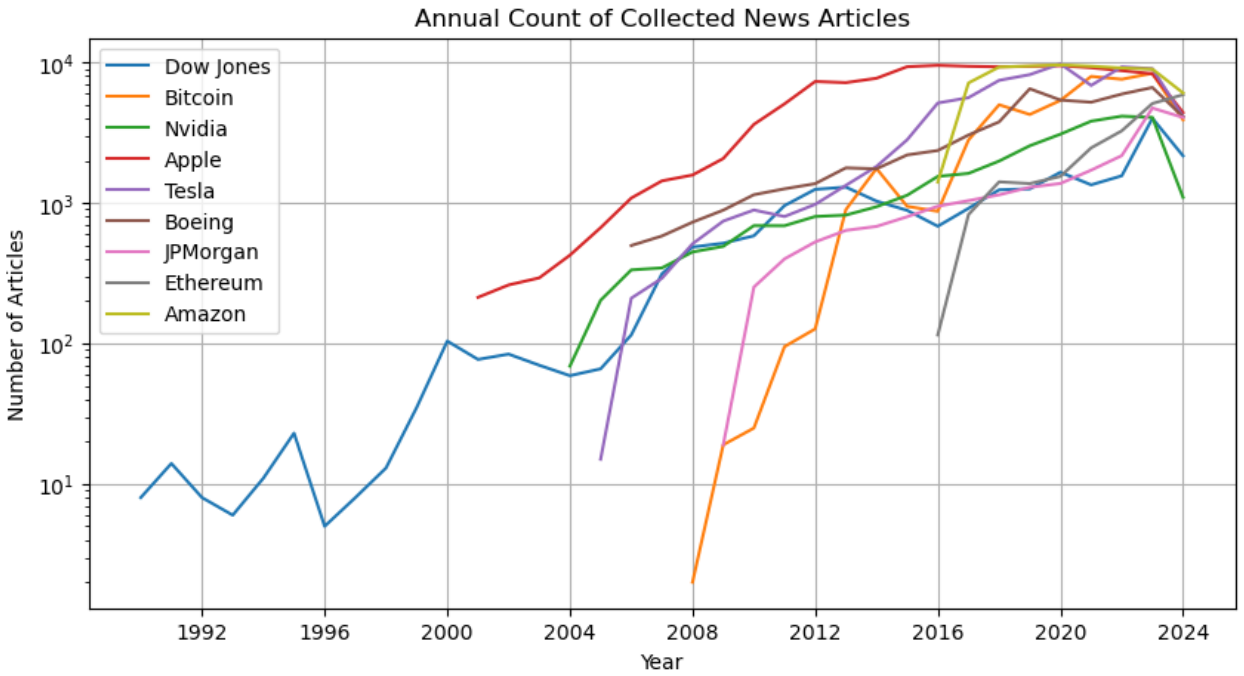

Figure 16: Yearly article coverage over time. The y-axis is on a logarithmic scale to show the variation in the number of articles across different datasets. Note that the number of articles in 2024 is lower because the year is not yet complete.

equal and includes many well-known outlets such as Business Insider, CNBC, AWS Blog, CNN, Yahoo Finance, and Bloomberg.

Table 1: Summary of data statistics. This table provides an overview of the dataset used in the analysis. These statistics help in understanding the scope and distribution of the data collected for sentiment analysis.

Statistic	**Value**
Total number of articles	483,691
Total number of unique sources	13,080
Mean number of articles per source	36.98
Median number of articles per source	3

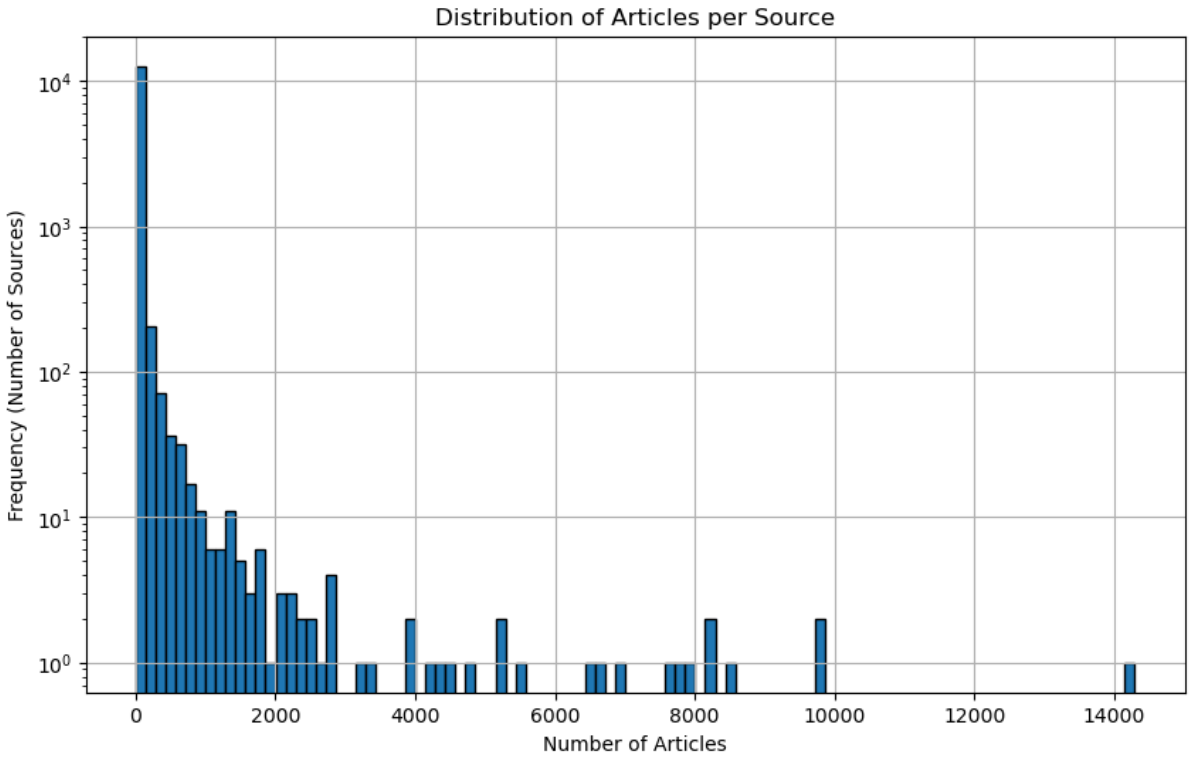

Figure 17: Distribution of articles collected per source. This histogram illustrates the frequency distribution of the number of articles collected per media outlet. The x-axis represents the number of articles, while the y-axis shows the frequency (number of media outlets). The leftmost bar shows that there are about 10,000 media outlets from which only 0-125 articles were collected. Conversely, the rightmost bar indicates that there is one media outlet from which roughly 15,000 articles were collected.

Finally, we might ask ourselves if the news articles contain any useful, forward-looking information, which has been shown to be the component of news with the most predictive power in the research of van Binsbergen et al. (2024). As illustrated in figure 19, this is often the case. Please note that the highlighted text in this figure is discussed afterwards.

For instance, the articles in our dataset frequently discuss upcoming market trends, stock predictions, and strategic decisions by companies that could influence future market movements. The first example article in figure 19 discusses the future of Nvidia's stock performance for the year 2024, which offers insights into market expectations and investment strategies based on current financial analyses and expert opinions. The second article explores the potential for Nvidia to split its stock, a move that could significantly impact investor sentiment and trading behaviors. This kind of forward-looking information is critical for investors seeking to anticipate market changes and adjust their portfolios accordingly. The final article from Business Insider discusses Nvidia's corporate strategies and employee compensation practices, providing a nuanced view of how these factors impact the company's long-term growth and market position. The article highlights the high compensation packages that Nvidia offers its employees but also examines the financial challenges faced

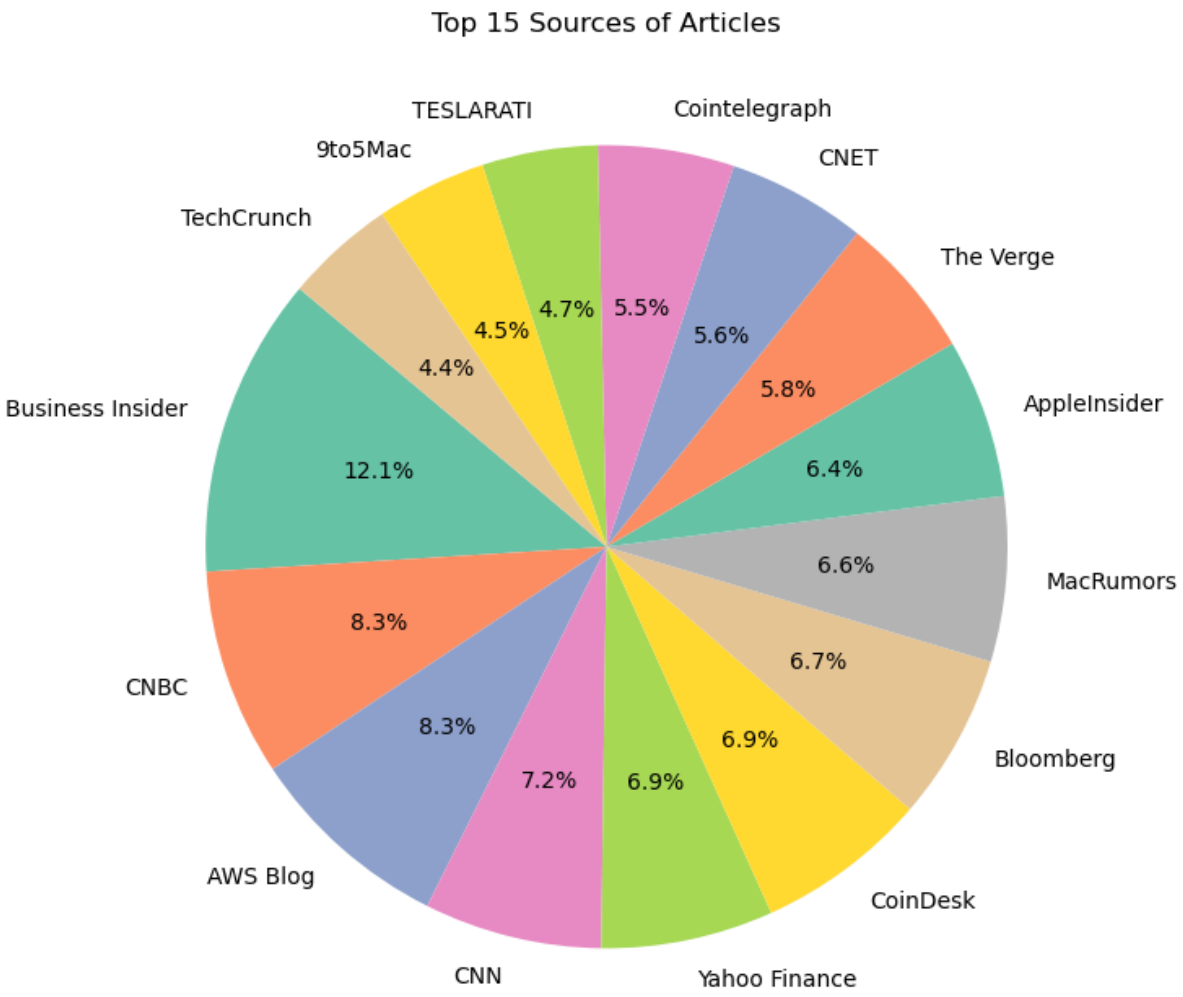

Figure 18: Article distribution among the top 15 news sources. This pie chart illustrates the proportion of articles contributed by the top 15 media outlets in the dataset. Each slice represents a media outlet, with the size of the slice indicating the percentage of total articles contributed by that source. While Business Insider is the largest contributor, the distribution among the remaining sources is relatively balanced.

by employees due to high living costs and taxes. These insights into employee compensation and corporate strategy can provide valuable context for understanding Nvidia's market position and potential future performance. Overall, we see how these articles present current news and also analyze potential future scenarios, making them useful for constructing a sentiment index with capabilities to help predict market trends.

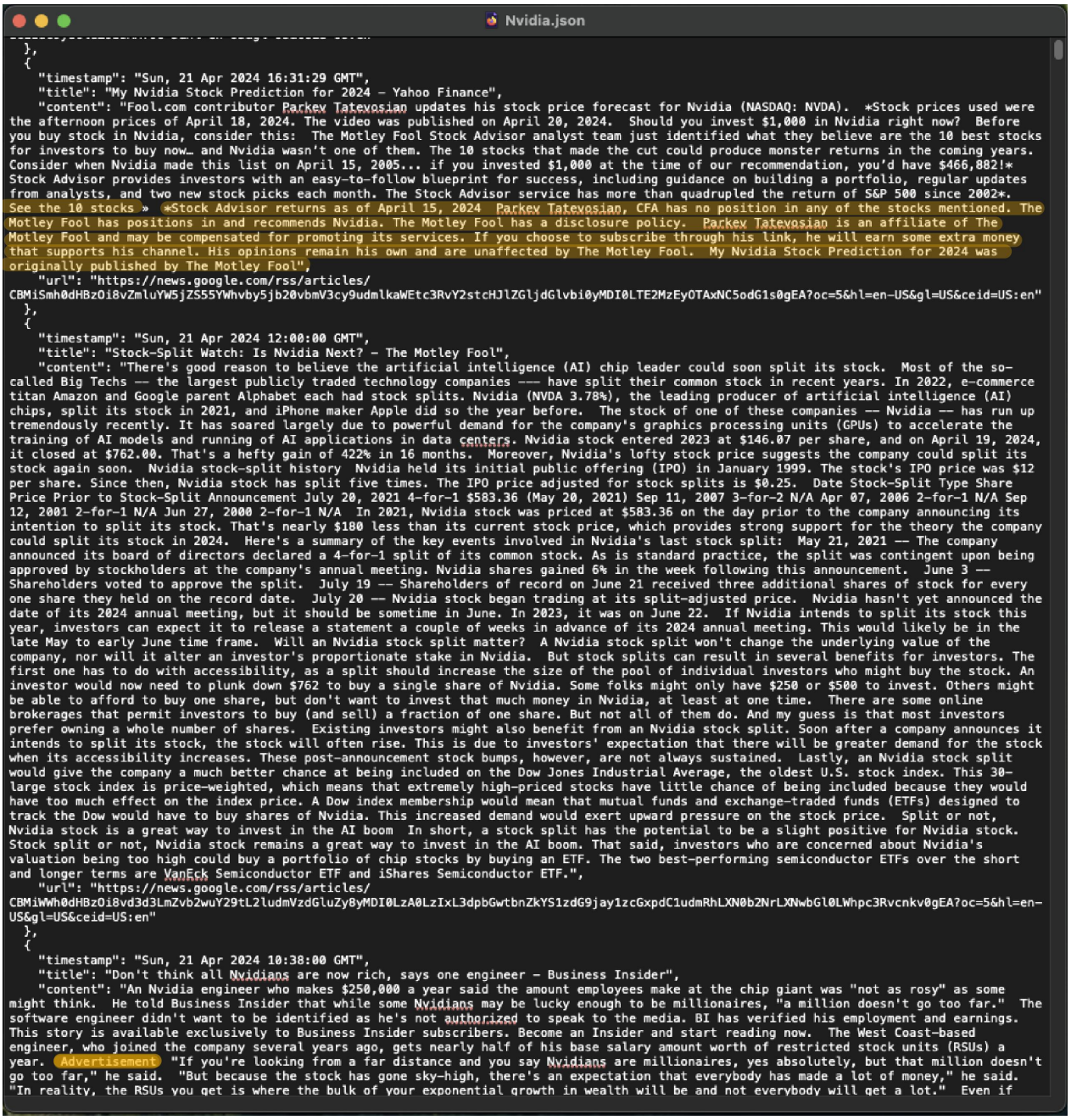

Figure 19: Example of processed data. This figure displays a snippet of the processed JSON data for articles related to Nvidia. The articles primarily discuss Nvidia's stock performance, potential stock splits, and market trends. However, even in this final data, some texts include irrelevant content such as disclaimers and promotional language. These elements highlight potential limitations for accurate sentiment analysis.

Before moving to the next subsection, where we discuss the chosen natural language processing (NLP) model and its implementation for classifying the sentiment of news articles, it is important to address several potential limitations of the data. Firstly, articles from many top-tier news outlets, such as The Wall Street Journal, Reuters, and Forbes, cannot be collected via the requests made by

GNews due to these outlets' measures to prevent automated news collection and/or their requirement for paid access. Additionally, as shown in Figure 20, some news outlets exhibit a lower percentage of negative articles, which may limit their contribution to a balanced sentiment analysis. Furthermore, as illustrated in Figure 16, Google News provides less data the further back in time one goes, leading to limitations in historical sentiment analysis. Finally, despite the data cleaning efforts, some articles still contain words or texts which are unrelated to the article, like the disclaimer and advertisement, highlighted in figure 19.

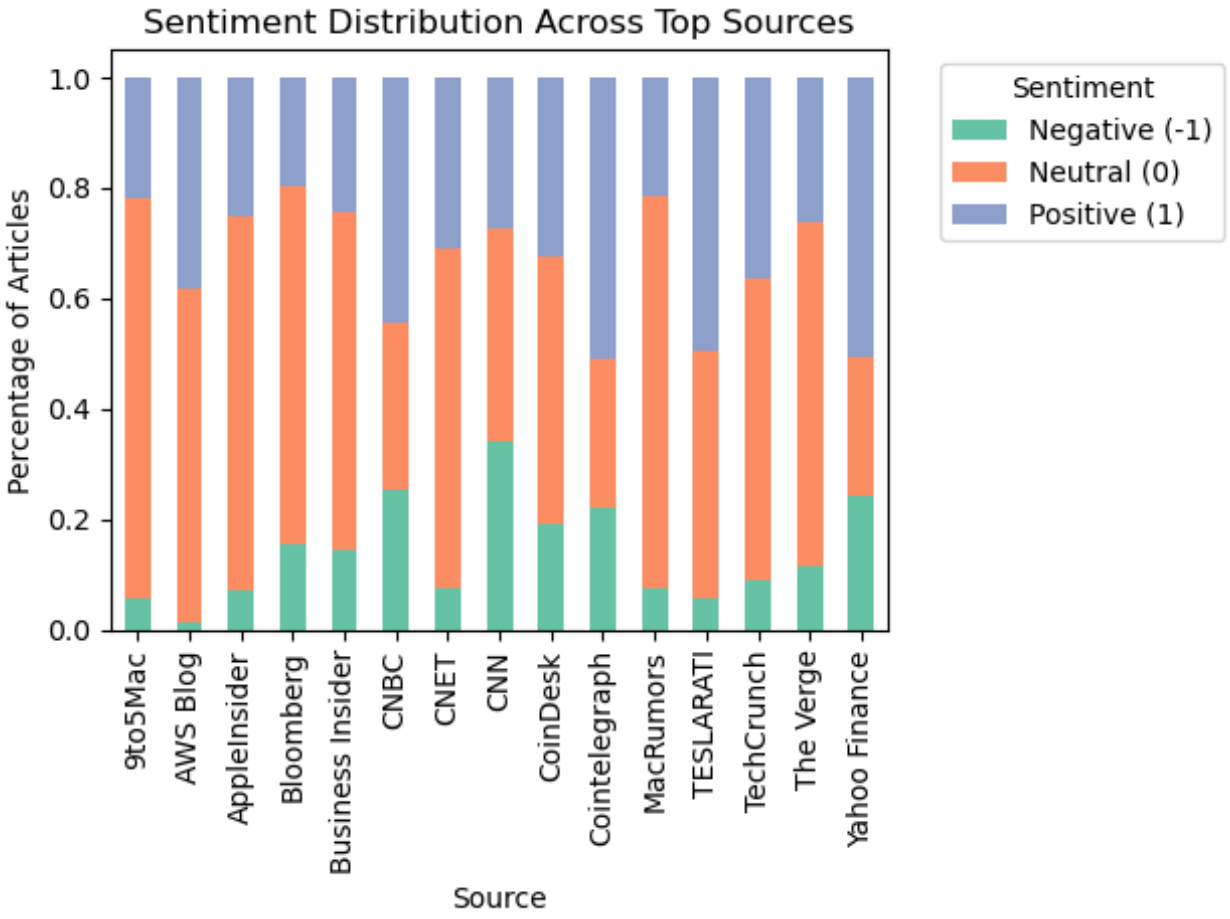

Figure 20: Illustration of the distribution of sentiment (negative, neutral, and positive) across the top 15 media sources in the dataset. Each bar represents a media outlet, showing the percentage of articles classified under each sentiment category. The chart highlights how sentiment is distributed differently among various sources, with some, such as CNN and CNBC, displaying a more balanced sentiment distribution, while others like AWS Blog and 9to5Mac exhibit a higher proportion of neutral or positive sentiment.

While these limitations might carry implications for the research, which are discussed in section 6, the collected data should still be considered both suitable and reliable for this research due to several reasons. First of all, Google News aggregates articles from a wide range of reputable sources. While the data used in this study excludes the mentioned premium outlets, it still includes content from

numerous respected news organizations and financial publications such as Yahoo Finance, CNN, CNBC, Bloomberg, and Business Insider, as illustrated in Figure 18. These sources are renowned for their journalistic standards and offer a broad spectrum of market-related news and analysis. The differences in sentiment distribution among these top sources do not inherently present problems for sentiment analysis. For example, the tendency of some outlets like 9to5Mac or AWS Blog to predominantly publish positive articles on certain topics does not imply that other sources will follow the same pattern. In fact, having outlets with varying sentiment distributions may enrich the dataset, as it captures a wider range of market sentiment, from optimism to pessimism. This diversity can be advantageous as it provides a more nuanced and comprehensive view of the market.

The limited availability of historical data is an unavoidable challenge, not unique to this research. As time passes, older articles become less relevant, yet storing them incurs the same costs as storing current articles. Particularly before the digital era, paper-based articles needed to be digitized, further increasing the cost of preserving historical data. While including data from the distant past could potentially lead to more holistic conclusions, it does not prevent us from addressing the primary research question: whether we can enhance sentiment analysis using NLP.

Furthermore, the noise present in some articles may not significantly impact sentiment classification as much as one might expect. Keeping in mind that the specific NLP model and sentiment classification method are detailed in the next subsection, Figure 21 demonstrates that, despite slight shifts in probabilities when advertisements are removed from the final article in Figure 19, the overall sentiment classification remains negative. Since the sentiment label with the highest probability determines the final outcome, the noise in articles introduced by advertisements is unlikely to alter this decision in most cases.

```
Sentiment analysis for text with advertisements:
Label: negative, Probability: 0.7432
Label: neutral, Probability: 0.1118
Label: positive, Probability: 0.1449
Final decision: negative

Sentiment analysis for text without advertisements:
Label: negative, Probability: 0.8483
Label: neutral, Probability: 0.0944
Label: positive, Probability: 0.0574
Final decision: negative
```

Figure 21: Impact of advertisements on sentiment classification. The results demonstrate how the models probabilities for each class changes, but that the final sentiment label (negative), doesn't change.

3.2 The Natural Language Processing Model

There exist a wide range of NLP models that are well-suited for sentiment classification of texts. However, considerations must be made regarding the type of data being classified, the computational costs of training, and the speed of classification of certain NLP models.

3.2.1 Model Selection

In this research, a pre-trained DistilRoBERTa model, fine-tuned for sentiment classification of financial news, was chosen from the Hugging Face Transformers library[5], and achieves an accuracy of 98.2%. For a popular task like this one, using pre-trained models offers an efficient solution without the need for extensive computational resources and data required for training. Training a custom model from scratch is often necessary only for very specific tasks where existing pre-trained models may not perform adequately.

DistilRoBERTa, a distilled version of the Robustly Optimized BERT (RoBERTa) model, balances accuracy and efficiency (Liu et al., 2019; Sanh et al., 2019). It retains much of the original RoBERTa model's accuracy while being faster and less resource-intensive, making it well-suited for applications where computational efficiency is crucial. Chen et al. (2022) also demonstrated that the RoBERTa model is among the best performing for news-based return predictability. In this study, the use of DistilRoBERTa significantly improves the efficiency of classifying nearly 500,000 articles compared to the standard BERT model. This choice allowed for high accuracy in sentiment classification while managing computational resources effectively, ensuring that the model could handle the large volume of data efficiently, and also quickly for use in real-time analysis.

3.2.2 Implementation

The first step requires loading the pre-trained model using the `AutoModelForSequenceClassification` and `AutoTokenizer` classes from the Hugging Face Transformers library in Python. Each article is independently tokenized to prepare it as input for the classification model. Tokenization involves breaking down the text into smaller units called tokens, which are then mapped to unique numerical identifiers that the model can process.

Based on an input text, the model outputs a set of three logits, which are raw prediction scores for the classes "Negative", "Neutral", and "Positive". Logits are the unnormalized scores that the model outputs, representing the relative likelihoods of each class before any probabilistic interpretation. Mathematically, we denote the output logits as $\mathbf{z} = [z_1, z_2, z_3]$, where z_i represents the logit corresponding to class i.

The logits $\mathbf{z}$ are then converted into probabilities using the softmax function (1), which normalizes the scores to a range between 0 and 1, making them interpretable as probabilities. The softmax function ensures that the sum of all probabilities equals 1, thus providing a clear and interpretable measure of confidence for each class.

[5] `https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis`

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{1}$$

The sentiment of a news article is determined by the class with the highest probability. For example, if the probability for the "Positive" class is higher than that for the "Negative" and "Neutral" classes, the article is classified as having a positive sentiment.

3.2.3 Sentiment Index Construction

In line with Baker and Wurgler (2007) we construct a change sentiment index and a level sentiment for all assets. Once all the articles have been classified, the methodology of Hiew et al. (2019) is followed to construct the sentiment changes index, where articles after market closing times are considered as sentiment data for the next trading day. Based on any keyword k and timeframe t, the sentiment value based on that keyword in the given timeframe is calculated as follows:

$$CI_t^k = \frac{Pos_t^k - Neg_t^k}{Pos_t^k + Neu_t^k + Neg_t^k} \tag{2}$$

where Pos_t^k, Neu_t^k, and Neg_t^k represent the number of positive, neutral, and negative articles, respectively, for keyword k in timeframe t.

Intuitively, this formula works as a weighted average: the numerator $(Pos_t^k - Neg_t^k)$ captures the net sentiment by assigning scores of 1 to positive articles, -1 to negative articles, and 0 to neutral articles. The denominator $(Pos_t^k + Neu_t^k + Neg_t^k)$ normalizes this by the total number of articles. This ensures the sentiment value is proportionate to the overall volume of sentiment. If there are many articles with strong positive sentiment, the index reflects this strongly, whereas if there are fewer articles or they are predominantly neutral, the index moderates accordingly.

The level indices are constructed by taking the cumulative sum of the standardized sentiment values in (2):

$$LI_t^k = \sum_{j=1}^{t} \mathrm{CI}_j^k$$

This is justified by several arguments:

Firstly, sentiment values based on short time frames like days or months can be volatile and subject to short-term noise. Cumulative summation helps to smooth out these fluctuations, highlighting long-term trends and reducing the impact of short-term anomalies. This enhances the ability to detect patterns and trends that might be obscured by daily or monthly variations.

Secondly, sentiment in financial markets often reflects not just isolated events but an accumulation of investor emotions and perceptions over time. By taking the cumulative sum, the ongoing build-up or depletion of sentiment is captured, providing a clearer picture of the overall mood.

The level indices, are again standardized for consistent comparison.

3.3 Evaluating the Sentiment Indices

Distinct strategies are utilized to demonstrate that the constructed sentiment indices contain economically relevant and significant information.

A first step in deriving this value, can be achieved through the visual comparison of the NLP-based Market Sentiment Index (MSI) and the Michigan Consumer Sentiment Index (MCSI). If the MSI exhibits similar trends and patterns as the MCSI, it presents a significant advantage over traditional methods, as unlike the MCSI, which incurs high maintenance costs and is updated only once a month, the MSI can be constructed at nearly zero cost and updated at short intervals. The MCSI data can be retrieved from the Federal Reserve Bank of St. Louis[6] in Python.

3.3.1 Cointegration and Correlation Analyses

Next, a correlation analysis is performed to quantify the relationship between indices. Correlation measures the strength and direction of the linear relationship between two variables. A high correlation between the MSI and MCSI indicates that the MSI is capturing similar sentiment dynamics as the MCSI (Shapiro et al., 2022).

To ensure the robustness of the correlations, the stationarity of each index is tested using the Augmented Dickey-Fuller (ADF) test. The ADF test checks for the presence of a unit root in a time series sample. A stationary time series has properties such as mean, variance, and autocorrelation that are constant over time. Stationarity is essential for reliable statistical analysis and modeling of time series data. If the sentiment indices are non-stationary (contain a unit root), the correlations observed may be spurious, meaning that the relationships detected between the indices and other variables (such as the MCSI) might be due to random chance rather than a true underlying relationship. By ensuring that the sentiment indices are stationary, the ADF test helps confirm that the correlations and other statistical relationships are robust and meaningful.

If indices turn out to be non-stationary, additional cointegration tests are performed between them. The cointegration test is necessary because it examines whether there is a statistical long-term equilibrium relationship between indices, despite their non-stationary (and possible spurious) behavior. Two or more non-stationary time series are cointegrated if a linear combination of them is stationary. This implies that, even though the individual series may wander over time, they share a common economic driver and move together in the long run. If a cointegration relationship exists, the correlation between the indices is valid because it reflects a stable, underlying connection rather than a spurious one. A cointegration relation ensures that the observed correlations are not merely coincidental but indicate a meaningful and enduring relationship between the variables, providing further validity to the constructed indices. Besides this, it gives additional insights into whether asset-specific sentiment indices capture distinct sentiment trends unique to that asset, rather than perhaps market-wide or sector-wide trends.

[6] `https://fred.stlouisfed.org/series/UMCSENT`

3.3.2 Vector Error Correction Model

A powerful modelling framework for analyzing cointegrated time series, is the Vector Error Correction Model (VECM). Besides providing statistical proof of a cointegration relationship, a Vector Error Correction Model (VECM) also models how time series adjust and interact over time, capturing both short-term dynamics and long-term equilibrium behaviour. The general formula is:

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{p-1} \Phi_i \Delta X_{t-i} + a_t$$

where:

- ΔX_t represents the differenced vector of the variables at time t.
- ΠX_{t-1} is the error correction term, with Π being the matrix that captures the long-term equilibrium relationships and X_{t-1} being the vector of the variables in levels at time $t-1$.
  - The Π matrix can be decomposed into α and β vectors such that $\Pi = \alpha\beta'$:
    - α (the adjustment coefficients) indicates the speed at which the variables adjust towards the long-term equilibrium.
    - β (the cointegration vectors) captures the long-term relationships between the variables.
- $\sum_{i=1}^{p-1} \Phi_i \Delta X_{t-i}$ represents the sum of the short-term dynamics, with Φ_i being the matrices of short-term coefficients for each lag i and ΔX_{t-i} being the differenced vectors of the variables at previous times.
- a_t is the vector of error terms at time t, representing the unexplained components of the model.

Given that the MCSI is only available on a monthly basis, I opted to use a parsimonious model with one lag (p=2). Besides investigating the cointegration relationship, it allows us to explore whether a one-month lag in sentiment of MSI or MCSI causally influences the other. Under this decision, the model in equation 3.3.2 can be simplified as follows:

$$\begin{bmatrix} \Delta \text{MSI}_t \\ \Delta \text{MCSI}_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} \text{MSI}_{t-1} \\ \text{MCSI}_{t-1} \end{bmatrix} + \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} \Delta \text{MSI}_{t-1} \\ \Delta \text{MCSI}_{t-1} \end{bmatrix} + \begin{bmatrix} a_{1,t} \\ a_{2,t} \end{bmatrix}$$

If the graphs of the MSI and MCSI align, we could expect to observe a significant cointegration relationship. The adjustment coefficients in this relationship will provide insights into how quickly each series responds to disequilibrium. If consumers, surveyed by the University of Michigan, base their opinions on news articles, we might expect to observe a significant coefficient for the lagged MSI values on current MCSI values. This would indicate that trends discussed in the news gradually influence consumer opinions over time. If news articles primarily discuss past event, instead of being

forward looking while the consumers base their opinions on current information, we might see that current values of the MSI are significantly predicted by the MCSI.

3.3.3 Fama-French Five-Factor Model Regressions

For stock-specific sentiment indices, direct comparison is challenging due to the lack of established sentiment indices for these assets. Instead, Fama-French Five-Factor (FF5) regressions are employed. FF5 regressions assess the explanatory power of the sentiment indices relative to stock returns and traditional risk factors. The FF5 model (Fama & French, 2015) is an extension of the Fama-French Three-Factor model Fama and French (1993), which aim to explain the returns of stocks based on several risk factors, which are generally considered to capture systematic risks.

Through the inclusion of daily sentiment changes of a stock into the model, we can to measure whether these indices provide additional information beyond what is captured by the FF5 factors. The regression model is specified as follows:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,1}(\text{MKT}_t - R_{f,t}) + \beta_{i,2}\text{SMB}_t + \beta_{i,3}\text{HML}_t + \beta_{i,4}\text{RMW}_t + \beta_{i,5}\text{CMA}_t + \beta_{i,6}\text{CI}_{i,t} + \epsilon_{i,t}$$

where:

- $R_{i,t}$ is the return of asset i at time t.
- $R_{f,t}$ is the risk-free rate at time t.
- α_i is the intercept for asset i.
- $\beta_{i,1}$ to $\beta_{i,5}$ are the coefficients for the FF5 factors:
  - **MKT**: Market risk premium, $(\text{MKT}_t - R_{f,t})$.
  - **SMB**: Small Minus Big, the size factor.
  - **HML**: High Minus Low, the value factor.
  - **RMW**: Robust Minus Weak, the profitability factor.
  - **CMA**: Conservative Minus Aggressive, the investment factor.
- $\beta_{i,6}$ is the coefficient for the sentiment change index, the sentiment factor.

This model allows us to test the hypothesis that the sentiment indices provide significant explanatory power for asset returns over and above the FF5 factors. The inclusion of $\text{CI}_{i,t}$ enables us to determine whether sentiment indices capture additional, perhaps idiosyncratic, risk components that are not accounted for by the systematic risk factors traditionally used in asset pricing models. By incorporating this news-based sentiment into the regression model, we can gain deeper insights into its impact and relationship with asset values, beyond the fundamental and systematic information captured by the factors. The daily factor data is collected from Kenneth R. French's Data Library[7].

[7] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library

3.3.4 Sentiment Based Trading

To further investigate the effectiveness of the constructed sentiment indices, a trading strategy based on the crossing of moving averages is employed. Specifically, this strategy uses the 3-day and 3-week moving averages of the sentiment index. In this context, an n-day moving average at day t is the average of the index over the last n days, smoothing out short-term fluctuations and highlighting longer-term trends.

The crossing of these moving averages is significant for trading because it indicates potential changes in market sentiment. When the short-term moving average (3-day) crosses above the long-term moving average (3-week), it means that the average sentiment in the past 3 days is higher than the average sentiment over the past 3 weeks. This indicates a rapid and strong increase in sentiment, suggesting that a significant amount of positive news has been released related to a specific keyword. Consequently, this suggests a buying opportunity in the corresponding asset related to the keyword. Conversely, when the short-term moving average crosses below the long-term moving average, it signals a negative shift in sentiment, suggesting a selling opportunity. This method helps to identify and act on changes in market trends based on sentiment dynamics captured by the constructed indices.

The trading signals are generated based on the following criteria:

- **Buy Signal**: A buy signal is generated when the 3-day moving average (short-term) of the sentiment index crosses above the 3-week moving average (long-term). Formally, a buy signal is defined as $MA_3(SI_t^k) > MA_{21}(SI_t^k)$.
- **Sell Signal**: A sell signal is generated when the 3-day moving average of the sentiment index crosses below the 3-week moving average. This event indicates a negative shift in market sentiment, suggesting an opportune moment to exit or reduce long positions. Formally, a sell signal is defined as $MA_3(SI_t^k) < MA_{21}(SI_t^k)$.

The profitability of the trading strategy is assessed relative to the benchmark buy-and-hold strategy (buying at the start and selling at the end of the evaluation period). Specifically, the following metrics are used to evaluate performance:

- **Cumulative Returns**: The total return generated by the trading strategy over the evaluation period.
- **Sharpe Ratio**: A measure of risk-adjusted returns, calculated as the ratio of the strategy's excess return to its standard deviation.
- **Maximum Drawdown**: The largest peak-to-trough decline in the value of the strategy, indicating the risk of significant losses.
- **Win Ratio**: The proportion of trades that are profitable. This is calculated as the number of winning trades divided by the total number of trades.

By analyzing these metrics, the robustness and predictive power of the sentiment indices can be validated. A successful trading strategy based on these indices would provide strong evidence that the constructed sentiment indices effectively capture market sentiment and can be used to predict market movements.

It is important to note that the analysis of profitability in this context does not account for trading costs. The primary objective is to demonstrate that the sentiment indices contain relevant economic information. This analysis aims to validate the predictive power and informational value of the sentiment indices, rather than to develop a practical trading strategy that necessarily leads to monetary gains. Therefore, the focus is on the underlying economic signals rather than the net financial outcome after trading costs.

4 Results

This section conducts a comprehensive investigation into the information and value encapsulated in the different sentiment indices constructed in this research. Initially, we explore whether NLP-based sentiment indices offer improvements over traditional methods by comparing the NLP-derived Market Sentiment Index (MSI) to the Michigan Consumer Sentiment Index (MCSI). Through various econometric tests, we identify both short-term and long-term trends captured by these news-based sentiment indices. Notably, the MSI and MCSI exhibit a statistically significant cointegration relationship, indicating they share similar long-term trends despite short-term differences. This suggests that the MSI effectively captures components relevant to economic sentiment. With the MSI being constructed at minimal cost, with extreme speed, and capable of being updated at near real-time intervals, it offers significant advantages over traditional methods like the MCSI.

Building on this, we further investigate whether asset-specific sentiment indices provide unique insights beyond general market sentiment indices. Our analysis demonstrates that these indices often capture distinct information and trends specific to the asset or its sector, potentially reflecting events such as operational challenges and major product launches.

Interestingly, we also observe intriguing similarities, such as increased comovement in sentiment during periods of market distress. Broader economic disruptions can simultaneously impact sentiment across different assets, leading to higher correlations among sentiment indices. The rich duality of these indices underscores their potential for asset-specific, sector-wide, and market-wide analyses. For example, a significant decrease in all sentiment indices might serve as an early warning of market risk or distress, while a notable decline in indices related to a specific sector could indicate sector-specific risk.

The Fama-French Five-Factor (FF5) regression results validate these observations, revealing that each asset-specific sentiment index significantly increases predictability in its own returns but generally not those of others. These findings furthermore provide insights into an asset's sensitivity to sentiment and its systematic and unsystematic risks. Contrary to previous research, market sentiment alone does not appear to effectively explain individual asset returns, highlighting the importance of considering these asset-specific sentiment dynamics for a more nuanced understanding of financial markets.

The significant return predictability suggests the potential for profitable trading strategies (Baker & Wurgler, 2007). We conclude this section by demonstrating that a simple trading strategy based solely on the Bitcoin sentiment index outperforms the benchmark buy-and-hold strategy across several key metrics. The seemingly superior performance of a strategy that is generally too simple to be effective using price data alone indicates that the sentiment data captures valuable information relevant to price movements. Although trading is not the primary focus of this research, this illustration underscores the potential utility of the sentiment indices developed, highlights the possible comovement between sentiment and prices, and motivates future research in this area.

4.1 The Level Indices

We begin by presenting both the level index of the MSI and the MCSI.

The level indices of the MSI and MCSI in Figure 22 offer insights into their long-term trends. Overall, we observe that both indices tend to follow a similar trajectory over time. Specifically, they show an increase during the period from 2013 to 2018, a decline from 2020 to 2022, and another upward trend beginning in 2023.

Furthermore, both indices exhibit a notable negative shock in sentiment around the COVID-19 pandemic, aligning with the intuitive understanding that global events impact sentiment. This concurrent decline highlights the similarity in sensitivity of these indices to macroeconomic conditions and major market disruptions.

These observations are further substantiated by the high correlation of 82% between the two level indices. This correlation is significantly higher than those reported in previous studies, such as Shapiro et al. (2022) and van Binsbergen et al. (2024), which found correlations of their market indices with the MCSI to be 58% and 72%, respectively.

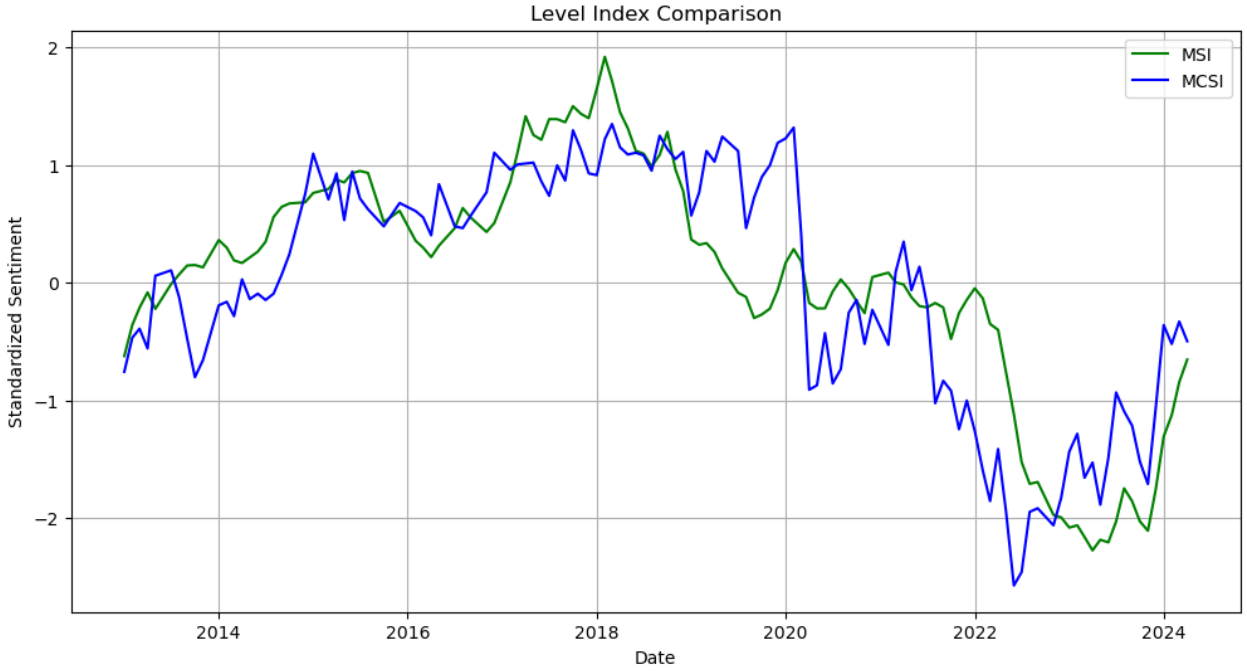

Figure 22: Level index of market sentiment index and michigan consumer sentiment index.

However, to avoid spurious conclusions, we must make sure that the series are stationary. Through visual inspections, we see some potential non-stationary behaviours of the indices: They seem to have trends over time, possible non constant variances and there are shocks visible in the series.

When performing Augmented Dickey-Fuller (ADF) tests, both level indices can not be concluded to be stationary. In fact, none of the level indices constructed for the other assets can be concluded to be stationary. Table 2 presents the ADF test statistics together with the corresponding p-values.

Table 2: Stationarity test results for all level indices

Index	**ADF Statistic**	**p-value**
MCSI	-1.84	0.36
MSI	-1.32	0.62
JPMorgan	-0.47	0.90
Amazon	-0.85	0.80
Apple	-1.83	0.37
Nvidia	-0.94	0.77
Boeing	-0.99	0.76
Tesla	-1.02	0.75
Bitcoin	-1.53	0.52
Ethereum	-0.49	0.89

Given the non-stationary nature of these indices, a cointegration analysis is a crucial next step in understanding the similarities and differences in the drivers of each asset's sentiment. Non-stationary series, when analyzed directly, often lead to spurious results, especially in regression models where trends in unrelated variables can produce misleading correlations. Cointegration tests allows us to identify pairs or groups of indices that, despite their individual non-stationary nature, move together over time due to a shared stochastic trend. By establishing these relationships, we can better understand the dynamics between different sentiment indices and how they may reflect common economic or market factors.

Table 3 presents the cointegration test results between the MSI and MCSI, and figure 23 presents a heatmap of the pairwise cointegration p- values between different sentiment indices.

Table 3: Cointegration test results between MSI and MCSI

Test	**Statistic**	**p-value**
Cointegration Test	-3.115	0.085

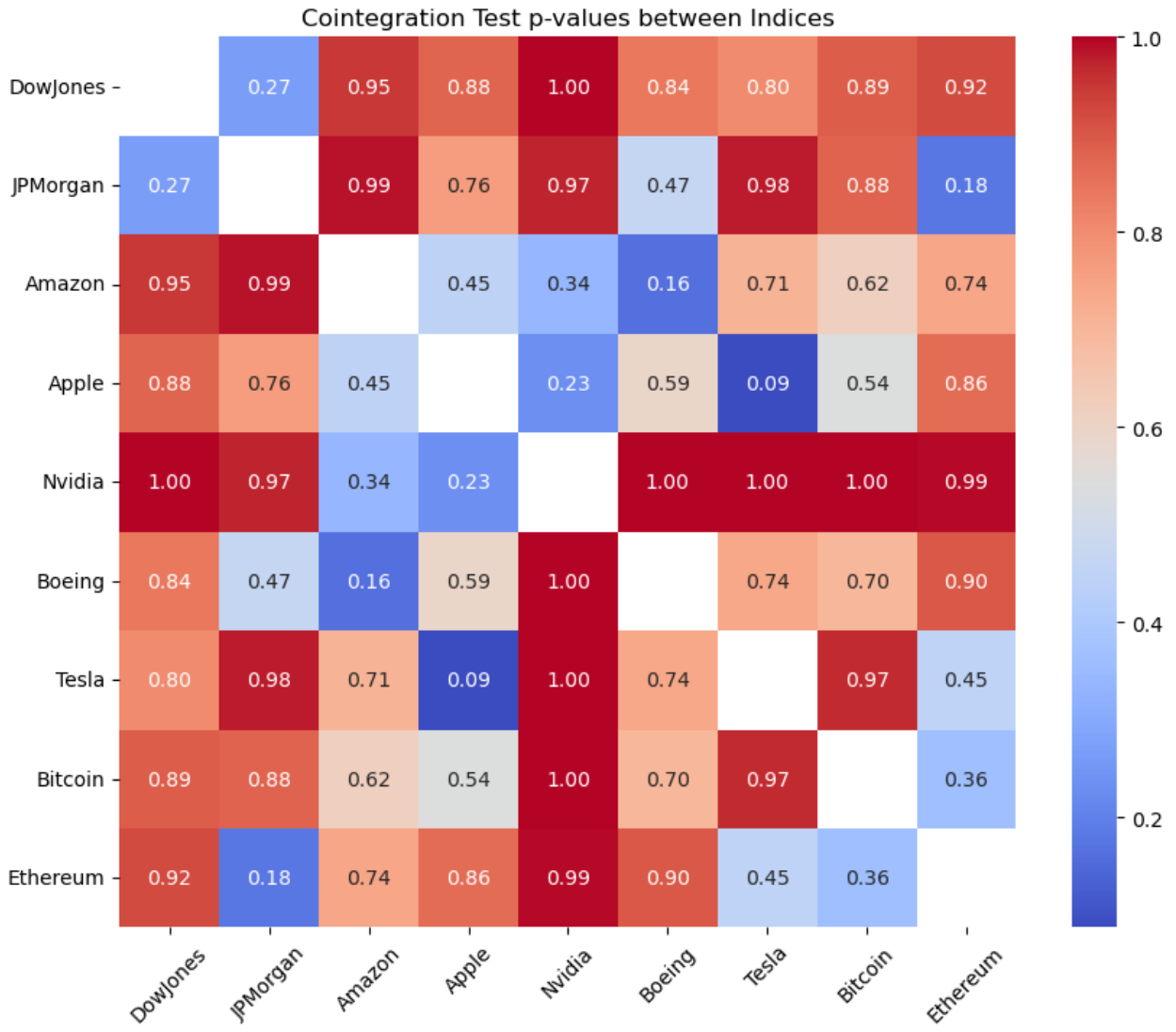

Figure 23: This heatmap illustrates the pairwise cointegration p-values between different sentiment indices. Lower p-values indicate a stronger likelihood of cointegration.

Some interesting observations can be made from these figures. First of all, most asset specific indices have very high p-values (as seen by the red colors), giving no statistical evidence for long term relationships between these series. However, in line with intuitive expectations, the p-values between tech-sector assets, cryptocurrencies, and the p-value between JPmorgan and the Market index are the lowest.

Intuitively, we would expect sentiment of assets in the same sector to have similar underlying components. The fact that we observe the lowest p-values between assets in the tech sector and cryptocurrency sector, supports this idea. Furthermore, as a major financial institution, JPMorgan plays a crucial role in the global financial system. Its operations reflect broader economic conditions, interest rate movements, and financial stability, and hence we would expect it to be most closely related to the market index compared to all other assets.

Perhaps most importantly, the lowest p-value is between the MSI and MCSI, which further validates the idea that our NLP market index is capturing common components to sentiment as the MCSI.

To further explore the relationship between the MSI and the MCSI, we fit a Vector Error Correction Model (VECM). The results in table 4 provide valuable insights into the dynamics between the NLP sentiment index and the Michigan sentiment index. The loading coefficient for the error correction term in the NLP sentiment index equation is marginally significant (p-value = 0.064), indicating a slow adjustment towards the long-run equilibrium. Similarly, the loading coefficient for the Michigan sentiment index is also marginally significant (p-value = 0.059), suggesting that both indices adjust towards the equilibrium over time, albeit gradually.

Table 4: VECM results

Variable	Coefficient	Std. Error	z-value
NLP Sentiment Index Equation			
L1.NLP sentiment index	0.3225***	0.087	3.717
L1.michigan sentiment index	0.0368	0.047	0.778
Michigan Sentiment Index Equation			
L1.NLP sentiment index	0.0487	0.181	0.270
L1.michigan sentiment index	0.0558	0.099	0.565
Loading Coefficients (Alpha)			
ec1 (NLP)	-0.0417*	0.022	-1.854
ec1 (Michigan)	0.0886*	0.047	1.891
Cointegration Relations (Beta)			
beta.2	-1.1343***	0.191	-5.937

*** p < 0.01, ** p < 0.05, * p < 0.1

The cointegration relation, as indicated by the significant beta coefficient (beta.2 = -1.1343, p-value = 0.000), confirms a long-term equilibrium relationship between the NLP sentiment index (MSI) and the Michigan sentiment index (MCSI). This implies that although the indices may diverge in the short term, they tend to move together in the long run, correcting any deviations from the equilibrium path.

Overall, these observations highlight that sentiment extracted from news seems to be a significant driver of economic sentiment. Despite some short-term differences between consumer sentiment and news-based sentiment, the long-term trends appear to be related. This reinforces the value of the MSI as both a distinct and complementary measure of sentiment trends, especially given that the MSI can be constructed free of charge and at near-continuous intervals, offering strong improvements over traditional measures like the MCSI.

Finally, we return to the question of what information each sentiment index is capturing and whether there are differences and similarities between them. The tests, graphs, figures and corresponding analyses allow us to answer these questions. In line with expectations, asset-specific

sentiment is influenced by asset-specific events, sector events, and market-wide events. This is highlighted by figure 24, which plots the sentiment of Boeing and Nvidia. According to the cointegration test results, these two assets have the least statistical support for any long term relations over time. Given that these companies operate in different sectors, we see distinct sentiment patterns that reflect their unique historical contexts. For example, the sentiment index for Boeing shows a significant decline starting around 2018, coinciding with the 737 MAX crisis and regulatory scrutiny, which heavily impacted its market perception. The recent plummet in sentiment starting in 2024 captures the impact of Boeing's ongoing safety and quality issues, which were highlighted by a major incident involving an Alaska Airlines 737 MAX 9 aircraft. The incident severely damaged Boeing's reputation, revealing potential lapses in manufacturing quality. Conversely, Nvidia's recent surge in sentiment likely reflects its growing leadership in AI technology and the success of its advanced chips, which are essential for meeting the rising global demand for AI applications.

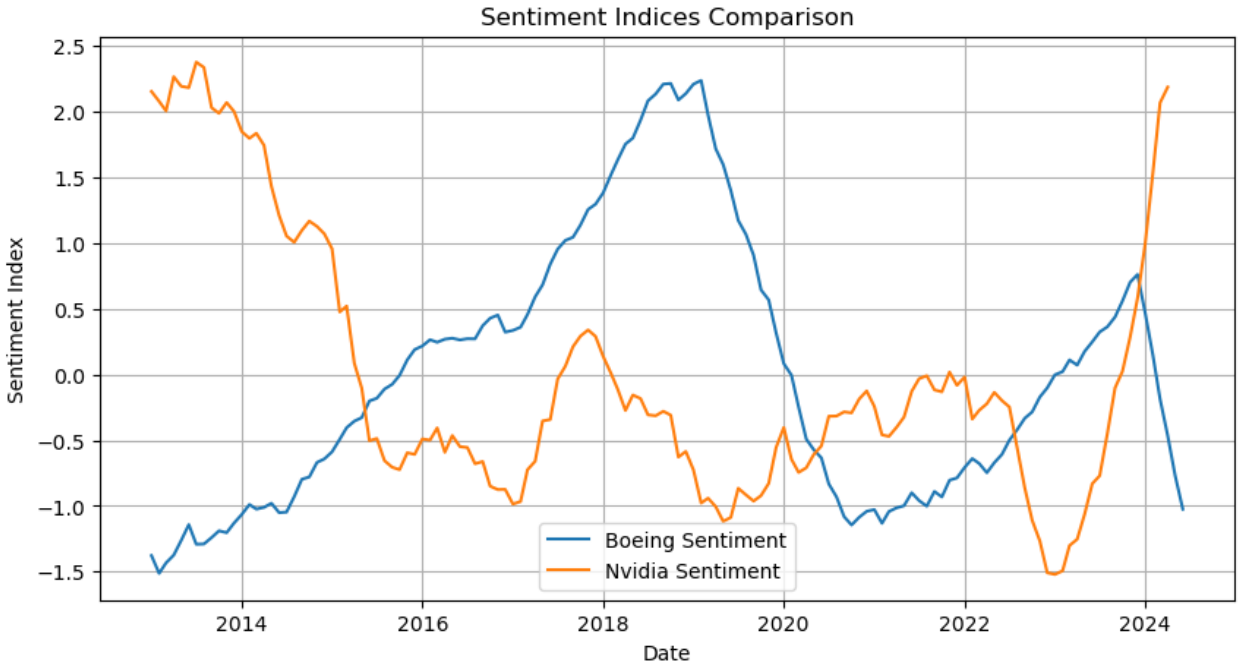

Figure 24: Sentiment of Boeing & Nvidia

Figure 25 illustrates the sentiment indices for Bitcoin and Ethereum. Unlike the previous comparison, these assets belong to the same sector, resulting in more noticeable similarities in their sentiment trends. Although some differences exist, both cryptocurrencies display remarkably similar patterns over time. For instance, the significant decline in sentiment during late 2017 and early 2018 likely corresponds to the bursting of the cryptocurrency bubble, when prices plummeted across all cryptocurrencies and market sentiment turned bearish due to regulatory concerns and market corrections.

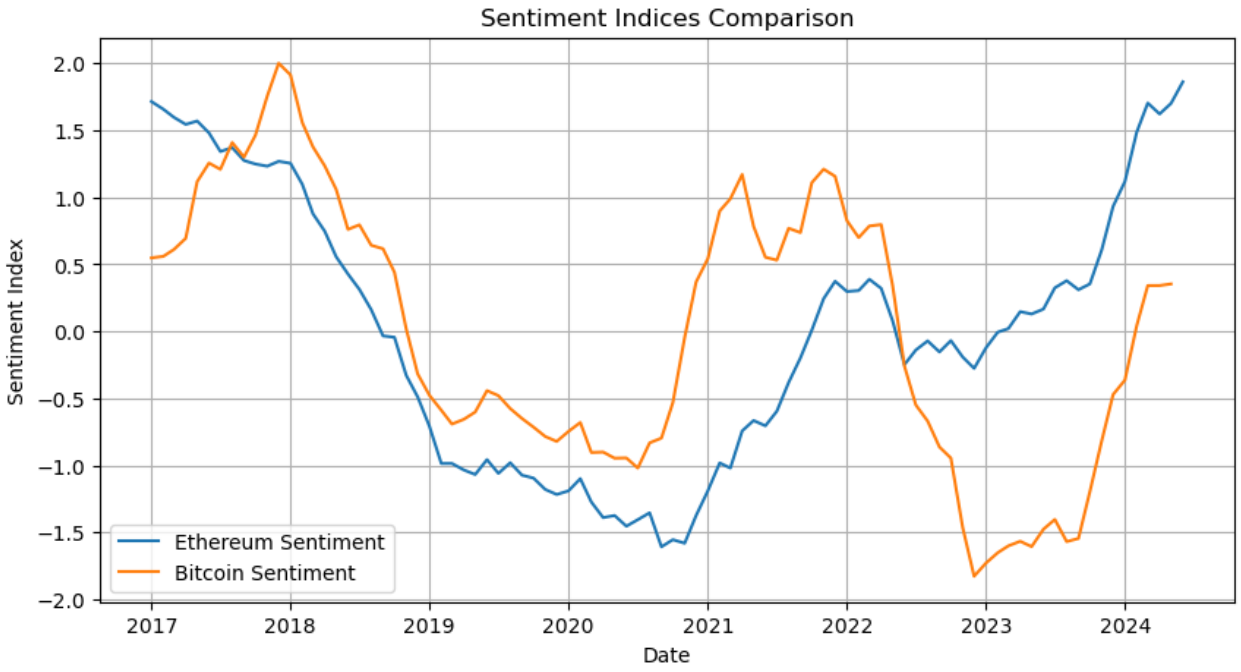

Figure 25: Sentiment of Bitcoin & Ethereum

Clearly, the sentiment indices effectively capture the impact of various types of events on an asset's sentiment. To further deepen our understanding of these indices, we examine the change indices, which represent short-term daily fluctuations in sentiment. This allows us to make more granular observations about short-term sentiment trends.

4.2 The Change Indices

Table 5 presents the ADF test results for all change indices, which are in essence the differenced series of the level indices. It is clear that all change indices are stationary.

Table 5: Stationarity test results for the change indices

Asset	ADF Statistic	p-value
MCSI	-9.9	0.00
MSI	-14.71	0.00
JPMorgan	-19.24	0.00
Amazon	-26.46	0.00
Apple	-38.28	0.00
Nvidia	-19.61	0.00
Boeing	-16.11	0.00
Tesla	-24.43	0.00
Bitcoin	-16.04	0.00
Ethereum	-15.04	0.00

To understand what the individual change indices are capturing, we focus on correlation analyses and visual inspections.

The change index of the MSI and MCSI is displayed in Figure 26. For better visibility, the range of the graph for the changes indices has been reduced. The changes indices exhibit a correlation of 31%, indicating a moderate positive relationship between the short-term movements of the MSI and MCSI. There are some observable similarities between the two indices, like at the end of 2023, and the negative shock in sentiment around the first months of 2020, capturing the COVID-19 crisis.

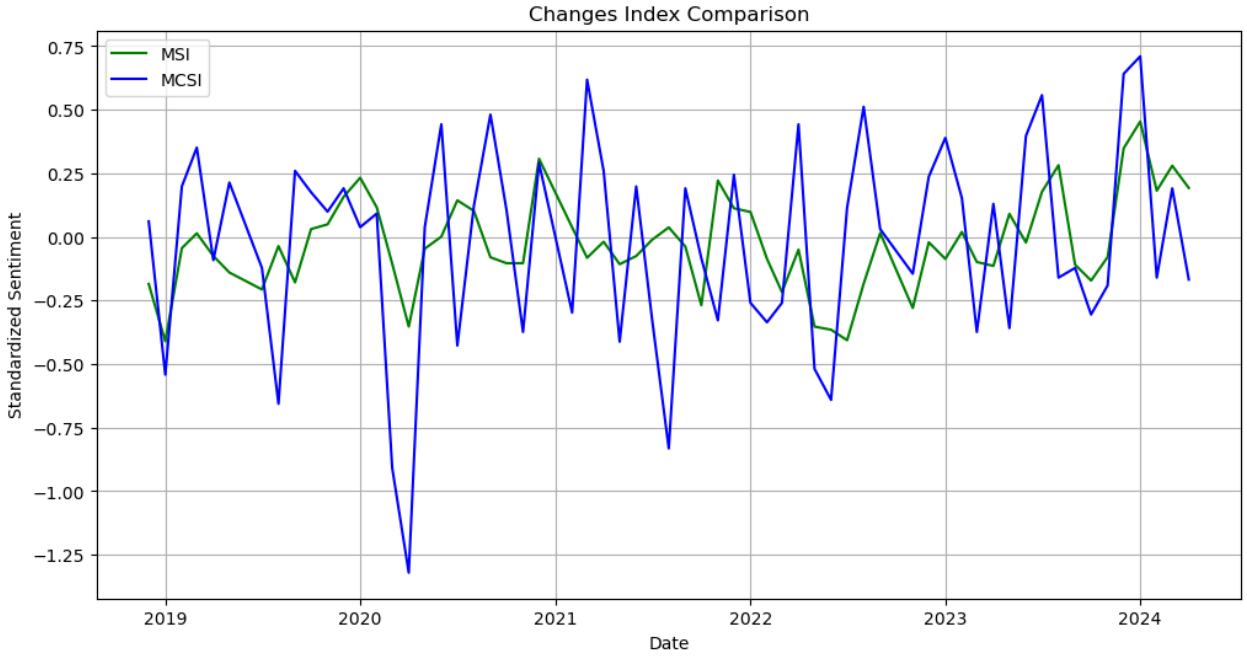

Figure 26: Index of sentiment changes for the MSI and MCSI

To get an idea of the degree of comovement in short term sentiment between all indices, a heatmap of correlations between all change indices is presented in figure 27.

Figure 27: Correlation heatmap. Each cell contains the correlation between two indices.

Most correlations are close to zero. The highest correlation, at 23%, is between the two cryptocurrencies, supporting the idea that short term sentiment trends between similar assets are driven by similar factors. The small correlations among all other assets suggest they capture different short-term information. Notably, the highest correlation of 31% between the changes indices is between the MSI and MCSI, further validating the idea that the market index captures components relevant to overall economic sentiment.

More granular insights can be derived from figures 28 and 29, which plot the 6 month rolling correlations between all change indices and all tech sector indices respectively. Before we discuss the main observations, please note that the correlation plots are only meant to demonstrate some

patterns which can be spotted in the sentiment index plots. These correlation plots provide a visual representation of how correlations between sentiment indices evolve over time and can help identify trends and periods of increased or decreased correlation. However, these rolling average correlations are by no means intended to serve as a formal econometric model for correlations, such as dynamic conditional correlation models or other advanced statistical methods. The primary purpose of these plots is to provide a preliminary, intuitive understanding of the relationships between sentiment indices, rather than a rigorous, dynamic measure of correlation.

First of all, we observe the largest increase in correlation between all indices during the COVID-19 Crisis, indicating that all sentiment indices were affected my this global market event. Correlations between sentiment in the tech sector show similar patterns, but most notably have much higher correlations compared to all indices together, reinforcing the idea that the sentiment of assets in a particular sector are driven by similar forces.

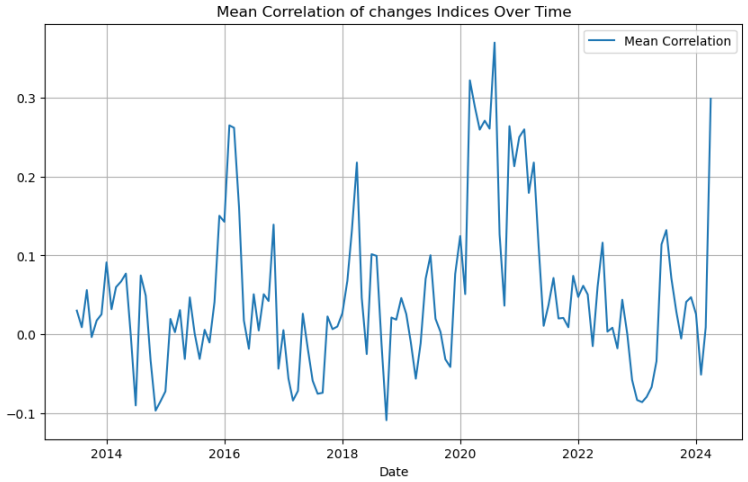

Figure 28: 6 month rolling correlation between all change indices.

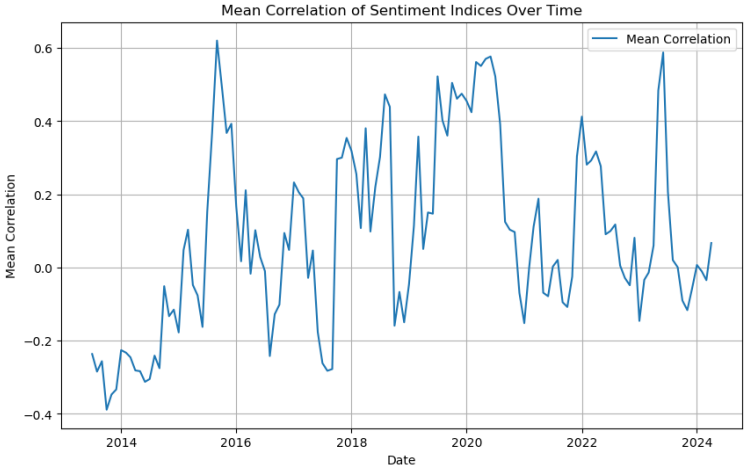

Figure 29: 6 month rolling correlation between tech sector change indices.

Our analyses of the level and change indices have deepened our understanding of the nuanced drivers of sentiment across different assets. We observed that each index captures unique sentiment trends specific to its respective asset, trends that are largely independent from those of other indices, as demonstrated by the correlation and cointegration results. However, events affecting specific sectors or the broader market tend to influence the sentiment of related assets in similar ways. This observation suggests a preliminary motivation for using changes in multiple NLP-based sentiment indices, which can be updated at near-continuous intervals, as a proxy for market risk. A simultaneous decline in sentiment across all indices could serve as an early warning sign of potential market instability.

The MSI's high correlation and cointegration relation with the MCSI highlighted the advantages of using NLP to construct sentiment indices, particularly in terms of minimal costs and rapid update capabilities. Given these benefits, a natural question arises: do the asset-specific indices offer similar advantages? Since asset-specific sentiment indices generally do not exist, and therefore lack established validation measures, we conducted various analyses to demonstrate their additional value, which are discussed in the following subsections.

4.3 Regression Results

Using 2,830 daily observations of trading days during the period from January 1, 2013, to April 1, 2024, I conducted a time series regression analysis using the Fama-French Five-Factor (FF5) model in Python. By incorporating daily values of the stationary asset-specific sentiment change indices into the model, we can assess whether these indices offer additional explanatory power over asset returns beyond what is captured by the FF5 factors, which are generally considered to represent systematic risks.

The regression equation thoroughly discussed in section 3.3.3, using Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors is:

$$R_{i,t} - R_{f,t} = \alpha_i + \boldsymbol{\beta}_i \cdot \mathbf{f}_t + \theta \cdot CI_{i,t} + \epsilon_{i,t} \tag{3}$$

Where:

- $R_{i,t}$: The return of asset i at day t.
- $R_{f,t}$: The risk free rate at day t.
- $\boldsymbol{\beta}_i$: Vector of coefficients for the five Fama-French factors for asset i.
- $\mathbf{f}_t$: Vector of the five Fama-French factors at day t.
- θ: Coefficient for the sentiment index $SI_{i,t}$.
- $CI_{i,t}$: Sentiment change index value for asset i at day t.

Table 6 presents the most important results, which show significant return predictability among all assets.

Table 6: FF5 main regression results (Using HAC standard errors)

Asset	CI Coefficient (θ)	Original Adj. R^2	% ↑Adj. R^2
Dow Jones	0.0001*	0.944	0.0
JP Morgan	0.0003*	0.764	0.0
Amazon	0.0008*	0.593	0.0
Apple	0.0016***	0.558	0.9
Nvidia	0.0012**	0.521	0.2
Boeing	0.0024***	0.434	1.98
Tesla	0.0056***	0.279	6.8
Ethereum	0.0088***	0.111	21.7
Bitcoin	0.0075***	0.059	45.8

*** p < 0.01, ** p < 0.05, * p < 0.1

Note: "Original Adj. Rsq" represents the adjusted R-squared value before including the sentiment index, indicating the proportion of variance explained by the original factor model. "% ↑Adj. Rsq" shows the percentage increase in the adjusted R-squared value after including the sentiment index.

The findings demonstrate that including sentiment indices significantly enhances the model's ability to explain variations in stock returns, suggesting that these indices capture unique information not fully accounted for by the Fama-French Five-Factor (FF5) model alone. By incorporating our earlier observations that sentiment indices capture idiosyncratic and unsystematic trends and news, while the FF5 factors account for more global, systematic events, it becomes evident that the sentiment indices improve return predictability by capturing risks that the FF5 model does not address.

The inclusion of market sentiment in the regressions yields minimal improvement in return predictability. For instance, in the case of Boeing, the coefficient for market sentiment is 0.0006 with a p-value of 0.107, and the adjusted R^2 shows only a 0.2% increase. Similarly, for Bitcoin, the market sentiment coefficient is -0.002 with a p-value of 0.030, indicating a statistically significant but small effect on return predictability. This is in line with van Binsbergen et al. (2024), who also find little evidence for return predictability using their market wide index.

4.4 Trading

As a final demonstration, we present a simple trading strategy, comparing its metrics to those of a benchmark buy-and-hold strategy. Additionally, we illustrate how the trading signals are generated.

From figure 30 and table 7, we observe that the sentiment-based trading strategy achieves higher cumulative returns than the benchmark. Furthermore, there is a significant decrease in maximum drawdown. While investors experienced a decline of more than 100% in the benchmark strategy,

the sentiment strategy only faced a maximum loss of 33%. The risk-adjusted returns are also significantly better, indicating that the returns from the sentiment strategy are considerably higher when accounting for risk.

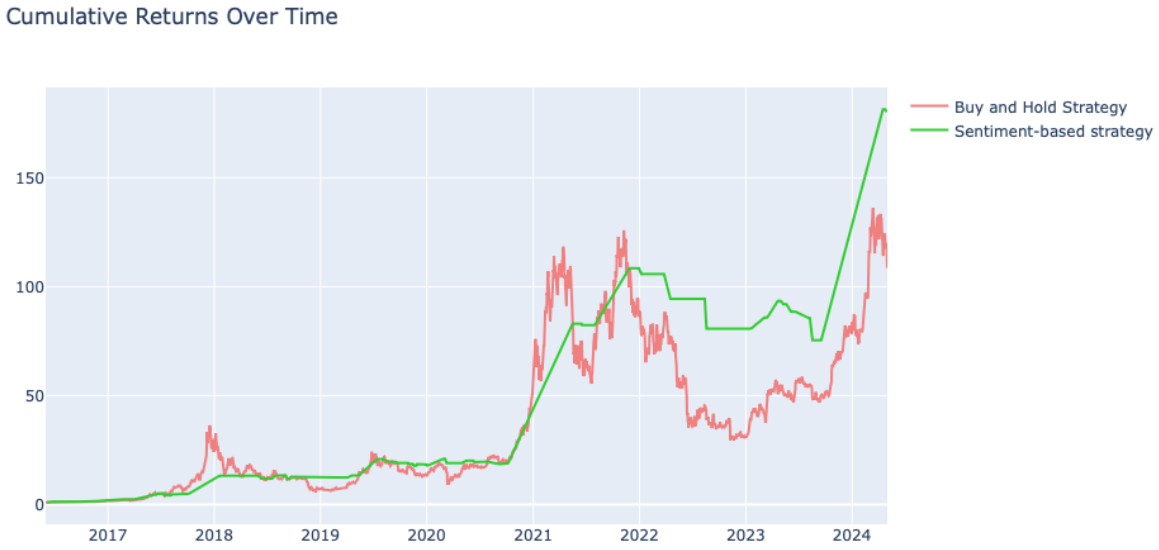

Figure 30: Comparison of trading strategies.

Metric	Sentiment Trading	Benchmark
Risk-Adjusted Returns	0.298	0.056
Win Ratio (%)	45.5	52.9
Maximum Drawdown (%)	-33	-113
Average Return (%)	16.5	0.2

Table 7: Performance metric comparison of trading strategies.

The win ratio and average return for the benchmark represent the average daily return and the ratio of positive to negative returns over the investment period. Interestingly, the sentiment strategy has a lower win ratio, with more trades yielding negative returns. However, the average return of all trades is 16.5%, suggesting that the trades with positive returns are of much higher magnitude than the losing trades.

To provide a clearer understanding of this trading strategy, in Figure 31 we plot the moving averages, showing how the buy signals (green) and sell signals (red) are generated. The corresponding plot displays the prices at which trades are made. Please note that the trading strategy in Figure 30 uses 3-day and 3-week moving averages. For clarity and visibility, longer-term averages were used in this demonstration.

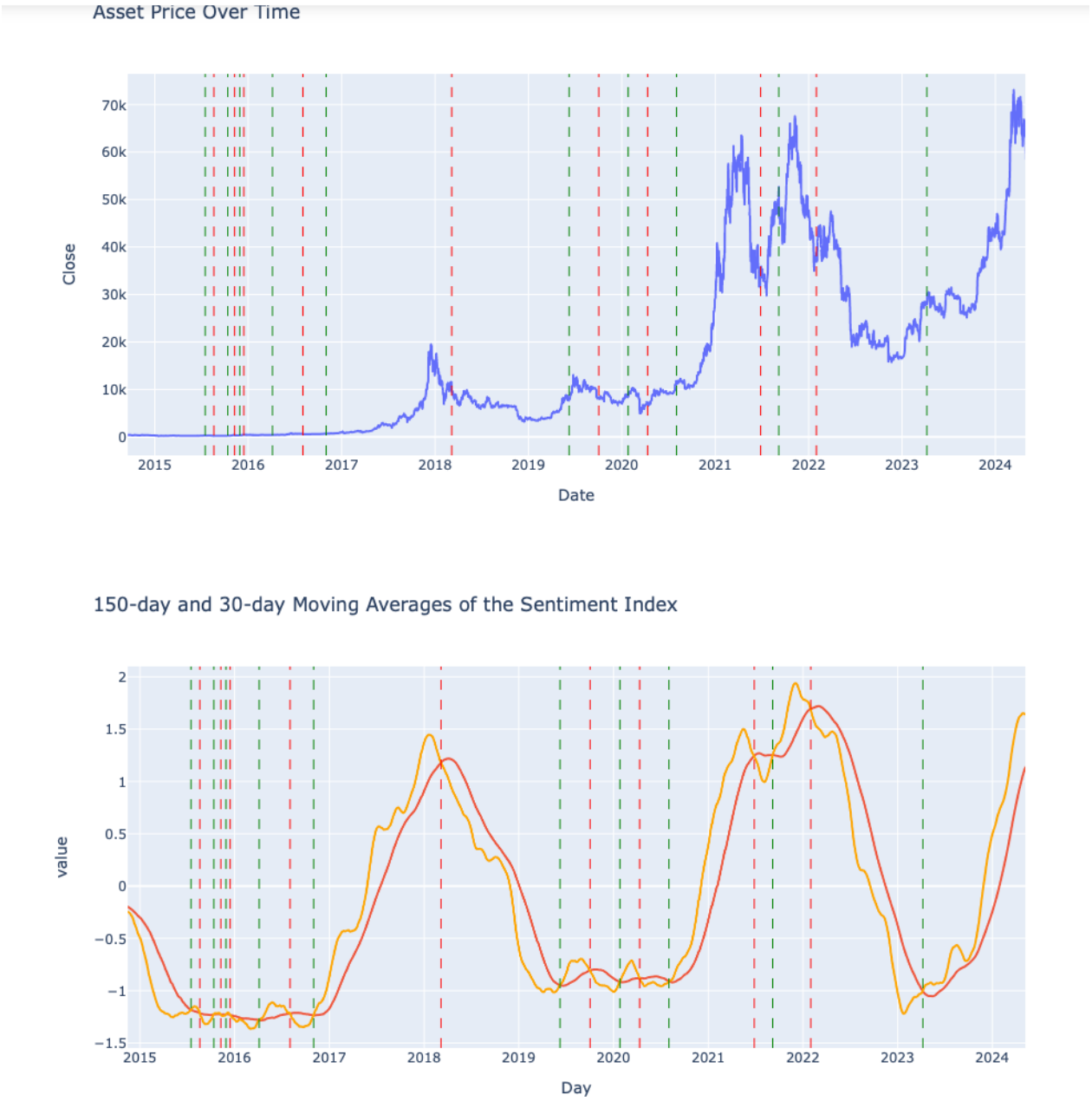

Figure 31: Trade prices with the corresponding generated trade signals. The green lines represent buy signals, while red lines are the sell signals.

In figure 31 we see how the trade signals are generated. We see some particularly profitable trades made in 2017-2018, 2021-2022 and 2023 onwards. The remaining shorter term trades are not as profitable, or even unprofitable and seem to be more noise driven. This is in line with the low win rate, but high average returns mentioned earlier.

Overall, the trading results are consistent with the return predictability results, and indicate that a sudden shock in news (as represented by the 3-day moving average) corresponds with trends in prices.

This strategy is by no means refined, but the fact that such a simple strategy leads to these above average results demonstrates the potential value contained in the sentiment indices. This motivates further exploration and refinement of trading strategies utilizing these indices, which will most definitely lead to even better outcomes.

5 Conclusion

This study set out to explore how natural language processing (NLP) could be used to enhance market sentiment analysis.

We have learned how traditional sentiment measures like the Michigan Consumer Sentiment Index (MCSI) are essential in understanding future economic trends like GDP growth (Curtin, 2004), even after accounting for the economic variables typically used in GDP forecasting (Howrey, 2001). Additionally, changes in sentiment were shown to precede recessions, giving strong economic motivations for policymakers and regulators to get hold of sentiment trends as early as possible. In financial markets, we have seen how sentiment is useful for investors, resulting in above-average profits (Tetlock, 2007), and in understanding the cost of capital for businesses (Baker & Wurgler, 2007).

These traditional works however, lacked the technology that we have today to effectively measure sentiment or relied on sentiment measures which were either costly and slow, restraining their value, especially during economic turning points (Shapiro et al., 2022).

This research directly addressed these issues by constructing low-cost, real-time sentiment measures, employing a custom-written script for free data collection, and using a pre-trained state-of-the-art NLP model, shown to be superior for sentiment classification of texts (Hiew et al., 2019).

Even though recent research has also already been applying advanced NLP technologies for their sentiment analyses (Chen et al., 2022; Shapiro et al., 2022; van Binsbergen et al., 2024), the focus often lies with market-wide analyses. One of the first studies using BERT to narrow the analysis to specific asset markets was Hiew et al. (2019). However, they don't explain the economic reasoning behind their results.

Besides extending market-wide analyses, with a focus on understanding the economic intuition, this work also concentrated on specific asset markets through the construction of a market sentiment index (MSI) and eight asset-specific sentiment indices. Using cointegration tests and a vector error correction model (VECM), we have learned how these asset-specific sentiment indices capture unique sentiment trends, whereas the MSI closely aligns with long term trends, and has the highest correlation in short term trends with the MCSI compared to all other asset index pairs. Furthermore we observed a cointegration relationship between our MSI and MCSI, with a high correlation of 82%. This correlation was higher than the work of Shapiro et al. (2022) and van Binsbergen et al. (2024), who obtained correlations of 58% and 72% respectively. Moreover, the asset-specific indices improved return predictions in Fama-French Five-Factor (FF5) model regressions, and were used to construct above-average profitable trading strategies, indicating that they provide additional value over regular market sentiment measures. Finally, we obtained insights into how specific indices might correlate in the presence of sector-specific or market-wide risks. Notably, we observed a dramatic increase in correlation among all stationary indices during the COVID-19 crisis, with indices related to the tech sector showing a consistently higher mean correlation.

Clearly, a lot of significant observations have been made. To properly grasp the implications of the findings, we first decipher the economic reasoning behind them, thereby strengthening the final

conclusions.

5.1 Giving Context to the Results

A natural first question that arises, is what the underlying relationship between the MSI and MCSI is, as indicated by the significant cointegration relation. The work of Hsu (2024), helps us answer this question, as it was found that news was an overall fundamental factor in forming consumer sentiment. The cointegration suggests that the MSI is effectively capturing a common component of the news which consumers are using to form their opinions and expectations. The VECM gave us more insight into the nature of this relationship. As hypothesized in section 3.3.1, the significance in lags of the MSI and MCSI in predicting their counterparts current value could tell us more about the causal relation between them. The insignificance of both lags in each other's equations doesn't prove, but suggests that both sentiment indices move contemporaneously based on the same information flow.

Moving from the market index, to the asset-specific indices, we might ask why these indices were so significant in improving return predictions, and constructing profitable trading strategies. First of all, through the works of De Bondt and Thaler (1985) and Tetlock (2007), we learned how the stock markets overreact to news. If the news contained no information relevant to the fundamental value of an asset, we would expect to see a complete correction of this sentiment-induced overreaction. As the FF5 regressions were conducted using closing prices, and mean sentiment changes computed at the end of the trading day, we would see no significant sentiment-induced return predictability if the overreactions in asset prices were corrected before the end of the trading day. Especially in large markets where many arbitrageurs are present, like that of Dow Jones, we would expect this to happen. The fact that we observe significant sentiment-induced return predictability among all assets, indicates that there is a component of the news captured which contains information relevant to the fundamental value of the asset. Additionally, the cointegration results validate the conclusion that all sentiment indices are capturing unique sentiment trends, unrelated to the overall market. Therefore, these indices are most likely capturing unsystematic, idiosyncratic news events, with a component of information that significantly influences asset prices. The favorable trading results might therefore stem from the fact that we are consistently trading on the most recent idiosyncratic news the moment that it is released. This gives us an informational advantage in financial markets as most regular traders can often not consistently be up-to-date with news the moment it is released.

Diving deeper into the variation in the results, specifically those of the FF5 regressions, we observe patterns that are in-line with earlier findings in the literature: the effects of sentiment in financial markets are exacerbated for smaller, harder-to-value assets (Baker & Wurgler, 2007; Chen et al., 2022; Tetlock, 2007).

Cryptocurrencies are commonly known to lack fundamental value indicators such as earnings reports or financial statements, making their value more subjective to speculation. Companies like Boeing and Tesla have among the lowest market capitalizations in the sample, and from the literature, we have learned how smaller stocks are disproportionately held by retail investors (Tetlock,

2007), who are commonly known to behave irrationally and be subject to sentiment. However, market size did not seem to be the only determinant. JPMorgan has a similar market capitalization to Tesla, but was far less sensitive to sentiment. Interestingly, we also observed how the original R-squared values in table 6 seem to be inversely related to the sensitivity to sentiment. Given that the factors in the FF5 model capture components of systematic risks, we can get an idea of an asset's returns comovement with overall systematic market conditions. Indeed, JPMorgan, being a financial institution, makes it significantly more exposed to certain types of systematic risks than Tesla due to its extensive involvement in various aspects of the global financial markets. This line of reasoning further validates the idea that the sentiment indices are primarily capturing unsystematic risks, and that this is the driver behind the observed results, as the assets which have a higher degree of systematic risk versus unsystematic risk have less sentiment-based return predictability. Hence, the sensitivity of an asset's return comovement with its sentiment index is likely due to to its sensitivity to idiosyncratic news. Another factor which is likely a component of the observed results, is limits to arbitrage (Shleifer & Vishny, 1997). Noise trader risk (De Long et al., 1990) in asset markets can prevent arbitrageurs from correcting the overreactions of investors to news. Especially in markets with predominantly retail investors, which are most subjective to irrational, sentiment based behaviour.

Finally, it is important to determine whether these results are temporary anomalies that will dissipate once enough investors catch on to them, or if they represent fundamental characteristics of financial markets. Research in the field of behavioral economics provides us with some insight into this question. Barberis and Thaler (2003) conclude that retail investors often fail to diversify their investments properly, and that the number of these investors entering financial markets is increasing. The implication is that many market participants are sensitive not only to systematic risks but also to unsystematic risks. Given that the improvements in FF5 return predictability likely stem from the indices' ability to capture a form of unsystematic risk through idiosyncratic news events, these results are likely to persist as long as investors continue to inadequately diversify their holdings.

Now that we have a better understanding of the economic interpretation behind the results, we can more effectively assess how these findings contribute to market sentiment analyses and the implications they bring.

5.2 Contributions and Implications

First of all, we have seen how the MSI offers significant improvements over the MCSI. Given that the MSI can be constructed at near-continuous intervals and most likely captures common components that the MCSI also does, regulators and policymakers have the ability to potentially track valuable sentiment trends, when the MCSI is not available. A sudden shock in the MSI, or among all asset-specific indices can give early warning signs of possible turbulence in specific sectors, or the entire market. For example, if there is a sudden shock in all sentiment indices relating to the financial sector, regulators can act quickly and take preventative measures to stabilize consumer sentiment, and reduce the impact of possible large scale financial disruptions. Governments could implement

temporary measures like trading halts or circuit breakers to prevent panic selling and maintain market stability. Regulatory bodies might issue reassuring statements to further calm markets and reduce panic, helping to stabilize sentiment among consumers and investors.

Second of all, the asset-specific indices further offer a valuable tool to both investors and businesses. Investors can create profitable trading strategies, and improve return predictions. The literature's suggestion that sentiment-induced return predictability are fundamental characteristics in markets motivates the use of these indices in constructing trading strategies, as they can deliver long term financial gains. Additionally, businesses can use asset-specific indices to understand trends regarding themselves, or the sector that they operate in. A sudden decrease in sentiment of their company will have implications for their future cost of capital and abilities to fund new projects, as investors might perceive them as increasingly risky, leading to sell-offs and lack of new investors. They can take timely action, come up with solutions to boost investor outlook of their company, or address the factors causing the decline in sentiment. For example, if a car dealership notices a rapid decrease in all sentiment indices related to the automobile sector, it can preemptively reduce prices or offer promotions to attract customers and slow down inventory orders, reducing the risk of overstock and minimizing potential losses while helping to maintain sales and market share if demand decreases significantly. Conversely, if it notes a sudden increase in Tesla sentiment, it can be the first to buy new cars before rising demand drives up Tesla prices, thereby increasing profits.

In conclusion, the main, all-encompassing value in these indices lie in their low-cost, timely nature, and capacity to compactly represent aggregated sentiment extracted from large volumes of textual data. Besides the ease of implementation and minimal costs making them accessible to all investors, businesses, regulators and policymakers, being able to immediately identify sentiment trends induced by the ability to process large volumes of text data in news articles offers a strategic advantage in multiple applications, and ultimately demonstrates the power that NLP gives us to enhance our sentiment analyses.

6 Limitations and Further Research

Having established the contributions of this research, it is essential to reflect on its limitations and suggest promising directions for future studies.

First, we consider the limitations and potential biases of the data. Given that most of the data was collected in the past decade, the analyses, results and hence conclusions are all based on the market conditions of this period. Additionally, data couldn't be collected from premium news outlets and many news outlets were more prone to releasing positive news articles than negative ones. It is therefore possible that the overall data used in this study was not as balanced and did not represent an accurate view of actual sentiment trends. The positivity bias might have made it less likely for the sentiment indices to capture negative trends in sentiment. Furthermore, the data collection methods through the custom-written script are reliant on Google News. It is possible that Google News implements changes to their service, or stops allowing automated article requests. In that case, the script must either be adjusted, or a different news distributor must be found. Finally, even after cleaning nearly five hundred thousand news articles, some article contents were unavoidably left with irrelevant texts like advertisements and disclaimers. Even though the observed results were still significant, and some preliminary investigation showed that the effects of noise in the data didn't significantly influence the sentiment classification, the exact effects are not completely understood, which could have further implications for the accuracy of the sentiment trends.

Next, we consider the NLP model. It is important to acknowledge that NLP models can learn significant biases through training data. If the financial news that it was trained on was not representative of average market news and sentiment, the model might have learned patterns in the data to classify sentiment that are not in line with average news, possibly leading to inaccurate classification of sentiment on our dataset and hence biasing the sentiment indices. Further considering that the NLP model was partially applied during times when it did not yet exist, it essentially implies that we were the only market participants with access to, and the ability to act upon, this technology. This is of course not representative of current market conditions, where increased competition among participants using NLP means that we will likely see a reduced window for profit and benefits in potential trading strategies using this technology. As a result, the observed results may not be as significant in the future as they were in the past.

Finally, only 8 asset-specific sentiment indices were constructed, which might not be completely representative of the broader market. While the literature suggests that the observations made in this research should at the least partially generalize to different market conditions and asset classes, there is no guarantee that we will observe the exact same effects moving forward.

Further research could immediately improve upon this study through the use of more balanced data, including premium news outlets and better data cleaning methods. Another suggestion might be to refine the constructed sentiment indices through weighting articles based on their overall relevance and credibility. This can lead to more accurate representations of sentiment trends and potential improvements in forecasting capabilities.

Another promising direction for further research is evident. We have seen how major studies

that relied on complex and costly models have demonstrated the potential of using NLP to construct news based sentiment measures for predicting economic fundamentals. The MSI used in this study showed a significantly higher correlation with the MCSI than those studies and can be constructed at minimal cost using a pre-trained model. Therefore, the potential payoffs for investigating whether the MSI can also be used in forecasting economic fundamentals are substantial.

Finally, as NLP technologies continue to advance, models are likely to become more sophisticated, not only in classifying the sentiment of articles but also in understanding more nuanced characteristics. For example, future models might be able to estimate the impact of information in a news article on asset prices. This research has shown that being an early adopter of such cutting-edge technologies can offer significant advantages. Therefore, it is recommended to explore and integrate these emerging technologies as soon as they become available to capitalize on their potential benefits.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–151.

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, *1*, 1053–1128.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1–8.

Chen, Y., Kelly, B. T., & Xiu, D. (2022). Expected returns and large language models. *SSRN 4416687*.

Curtin, R. T. (2004). Psychology and macroeconomics. *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, 131–155.

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793–805.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. doi: 10.1016/j.ejor.2017.11.054

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). BERT-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

Howrey, E. P. (2001). The predictive power of the index of consumer sentiment. *Brookings Papers on Economic Activity*, *2001*(1), 175–207.

Hsu, J. (2024). *Sources of Economic News and Information for Consumers* (Tech. Rep.). University of Michigan Surveys of Consumers. Retrieved from `https://data.sca.isr.umich.edu/fetchdoc.php?docid=75635` (Accessed: 2024-08-10)

Kelly, B., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends® in Finance*, *13*(3-4), 205–363.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1–19.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, *228*(2), 221–243. doi: 10.24148/wp2017-01

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168.

van Binsbergen, J. H., Bryzgalova, S., Mukhopadhyay, M., & Sharma, V. (2024). *(Almost) 200 Years of News-Based Economic Sentiment* (Tech. Rep.). National Bureau of Economic Research. doi: 10.3386/w32026

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73.