A Hype-Adjusted Probability Measure for NLP Stock Return Forecasting

Zheng Cao[*]

zcao26@jh.edu

Hélyette Geman[†]

hgeman1@jhu.edu

Abstract

This article introduces a Hype-Adjusted Probability Measure in the context of a new Natural Language Processing (NLP) approach for stock return and volatility forecasting. A novel sentiment score equation is proposed to represent the impact of intraday news on forecasting next-period stock return and volatility for selected U.S. semiconductor tickers, a very vibrant industry sector. This work improves the forecast accuracy by addressing news bias, memory, and weight, and incorporating shifts in sentiment direction. More importantly, it extends the use of the remarkable tool of change of Probability Measure developed in the finance of Asset Pricing to NLP forecasting by constructing a Hype-Adjusted Probability Measure, obtained from a redistribution of the weights in the probability space, meant to correct for excessive or insufficient news.

Keywords: Hype-adjusted Probability Measure, Natural Language Processing, Sentiment Analysis, Market Volatility Forecast, Semiconductor Industry

---

[*]Graduate Student, Johns Hopkins University

[†]Research Professor, Johns Hopkins University

1 Introduction

Sentiment has long been recognized as a key driver in financial markets, influencing both trading behavior and asset prices. Previous studies have examined the theoretical and empirical role of sentiment in market dynamics, highlighting its impact on investors' decision-making and trading patterns [12, 17]. Furthermore, research [2] by Baker and Wurgler (2006) followed by others has shown a significant relationship between investor sentiment and stock returns, providing a foundation for sentiment-based trading strategies .

Our manuscript builds on the work of Deveikyte et al [6], who applied Latent Dirichlet Allocation to forecast market prices and volatility. The authors proposed a thorough approach to compute the sentiment score and forecast the directions of the next day's market return and volatility in the case of FTSE100 stocks. In this research, we choose semiconductors, a very vibrant industry of this decade, as the sector for analyzing news and market data.

After a review of the literature on NLP and Changes of Measure in section 2, a detailed overview of data and methodology is presented in section 3, which includes the developments from previous sentiment equation to more sophisticated ones along with enhanced NLP forecasting methods. Building upon these foundations, we introduce in section 4, a novel probability measure, that we name " Hype-Adjusted Probability Measure", designed to capture the occurrence of market "hype." The results and future work are discussed in section 5.

2 Literature Review

Sentiment analysis, a subfield of Natural Language Processing (NLP), focuses on quantifying the emotional tone and intent conveyed in textual data. NLP itself is a rapidly growing area of machine learning (ML) that enables computers to process and understand human language through algorithms and statistical models such as the founding work by Jurafsky and Martin (2000) [14]. Within NLP, sentiment analysis applies techniques to assess positive, negative, or neutral sentiment in data, often used for applications like market forecasting and consumer behavior analysis [15].

In this context, sentiment is defined as the emotional polarity associated with a piece of text, often derived from news articles, tweets, or other media. It is typically measured using computational methods such as dictionary-based approaches or ML models. For example, Vader, a lexicon-based sentiment analysis tool, evaluates sentiment scores ranging from extremely negative to extremely positive [13].

In their 2022 paper, Deveikyte et al identified a significant negative correlation between daily negative sentiment and observed market volatility and develop a sentiment-based model for predicting directional volatility and market returns [6]. Building on their findings, we aimed to enhance the predictive relationship between news sentiment and stock performance by refining the methods used to calculate key financial metrics.

The return on day t are classically defined as the logarithmic change in the closing price from day $t-1$, expressed as:

$$r_t = \log \frac{\text{CLOSE}_t}{\text{CLOSE}_{t-1}} \tag{1}$$

The annualized volatility, σ, is calculated by the formula:

$$\sigma = \sqrt{\frac{1}{N} \sum_{t+1}^{N} (r_t - \bar{r})^2 \cdot \sqrt{252}} \tag{2}$$

Notice, for the programming and ML components of the manuscript, we selected a rolling window of 5 days to account for the amount of a regular trading week (without holiday breaks).

These metrics serve as the foundation for analyzing the connection between sentiment data and stock market behavior.

We adopted the Vader Sentiment engine to help translate multi-language texts into scores. It classifies sentiment scores on a scale as introduced by Clayton J. Hutto and Eric Gilbert [13]:

1. Extremely negative: -4

2. Neutral: 0

3. Extremely positive: +4

The sentiment score has been defined by many authors, in particular the paper [7] as the fraction of the difference between the number of positive and negative sentiment tokens, scaled by the sum of positive, neutral, and adjusted negative cases, of a given time d:

$$Sent_d = \frac{N_d(\text{pos}) - N_d(\text{neg})}{N_d(\text{pos}) + N_d(\text{neutral}) - N_d(\text{neg}) + 3} \tag{3}$$

Glasserman and Mamaysky (2017) highlight the predictive power of unusual news patterns in forecasting market stress [9]. Shapiro et al (2022) developed a novel sentiment-scoring model tailored to economic news articles, showing that daily news sentiment predicts consumer sentiment movements and macroeconomic responses to sentiment shocks, such as increased consumption, output, and interest rates [16]. Cohen et al. (2023) demonstrate the robustness of

multimodal classifiers under perturbation attacks and present the potential for integrating resilient, multimodal approaches into financial sentiment analysis for improved market forecasting [5].

Regarding the change of probability measures, it has been revealed as a remarkable tool in the finance of Asset Pricing. It started with the seminal paper by Harrison and Kreps (1979) [11], who proposed under No Arbitrage and constant interest rates a new probability measure that incorporates the risk premium attached to equities and leads to martingales for discounted stock prices. The "forward measure" introduced in Geman (1989) [8] was constructed to address the challenges created by stochastic interest rates in the valuation of risky cash flows.

To our best knowledge, no paper has yet introduced a change of probability measure in the context of NLP forecasting of asset returns and volatility. The instrument is beautiful, the application here original. Note that in the context of product management, Wind and Mahajan (1987) emphasized the value of "marketing hype" when launching new products in order to create a supportive market environment.

3 Data and Methods

As in the foundational paper of Bengio et al (2003) [4], we define in this paper 'news' as textual data, extracted from verified financial sources reporting on market trends, corporate developments, and stock performance. Primary data sources for this study were obtained from LSEG and the Eikon API [10], which provide access to over half a million news articles related to 30 semiconductor stocks. These data sources offer a comprehensive coverage of market events, corporate news, and sentiment indicators.

To mitigate potential biases in the collected data, particularly those arising from media overrepresentation or underrepresentation of specific events, we implemented two primary adjustments:

1. Assigning Weights to News Sources: Each news source is assigned a weight to achieve a more objective sentiment score adjustment and the adjusted score is calculated as:

$$\text{sentiment}_{\text{adj}} = \alpha \cdot \text{sentiment} + \beta \tag{4}$$

   where α represents the weighting factor applied to a news source, and β corrects for inherent biases.

2. Adjusting for Over- or Under-Reported Events: To address the uneven representation of events in the dataset, we use two key methods:

   - Removal of duplicate or near-identical articles to prevent overrepresentation of certain events.
   - Increase of the impact of under-reported events by assigning higher weights explained below.

3.1 Data and LDA Model Results

Appendix A presents a sample table of 30 tickers selected from the iShares Semiconductor ETF (SOXX). Note this ticker weight table is adjusted by removing 2 values: CME E-MINI S&P500-technology sector index future for September 2024, 0.15247% and CME Index and Options Market E-MINI Russell 2000 index future for September 2024, 0.15247%. The reason is that these two entries do not represent any actual company, but a mix of companies that are already in our list.

We first assemble the results using the previous sentiment methods on the semiconductor data:

1. the "daily weighted average" sentiment score: computed by the aggregated news data based on individual tickers, and adjust the processed sentiment score by individual tickers' weight from the overall SOXX sector (excluding the 2 Future ETFs removed)

2. the "overall daily average" sentiment score: searched by ticker name, without being adjusted by individual weights

3. the "overall daily average semi title" sentiment score: searched by the topic of SOXX sector, without being adjusted by individual weights

Each data set generates a accuracy precision report through a simple Linear Discriminant Analysis algorithm [3]. Linear Discriminant Analysis (LDA) is a commonly used dimensionality reduction technique that is effective for classification. It projects data onto a lower-dimensional space while maximizing separation between classes by modeling inter-class differences using the mean and variance within each class. Unlike Principal Component Analysis (PCA), which

maximizes variance, LDA prioritizes class separability, making it especially useful in supervised learning with labeled data.

Tables 1 2 3 below present the forecast results based on the simple LDA model, with volatility direction as the predicted target and sentiment scores as the only input training parameters.

For the "daily weighted average" sentiment score:

Class	Precision	Recall	F1-Score	Support
0	0.65	0.73	0.69	33
1	0.62	0.54	0.58	28
Accuracy			0.64	
Macro Avg	0.64	0.63	0.63	61
Weighted Avg	0.64	0.64	0.64	61

Table 1: Classification Report for Best Volatility Direction Model

For the second, the "overall daily average" sentiment score:

Class	Precision	Recall	F1-Score	Support
0	0.58	0.91	0.71	32
1	0.73	0.28	0.40	29
Accuracy			0.61	
Macro Avg	0.65	0.59	0.55	61
Weighted Avg	0.65	0.61	0.56	61

Table 2: Classification Report for Best Volatility Direction Model

For the third data input, the "overall daily average title" sentiment score, the result is

Class	Precision	Recall	F1-Score	Support
0	0.66	0.87	0.75	31
1	0.80	0.53	0.64	30
Accuracy			0.70	
Macro Avg	0.73	0.70	0.70	61
Weighted Avg	0.73	0.70	0.70	61

Table 3: Classification Report for Best Volatility Direction Model

The latter sections present a more sophisticated method of processing the sentiment score, with supporting results provided.

Figure 1 presents a sample distribution of calculated sentiment scores, showing a notable negative skew.
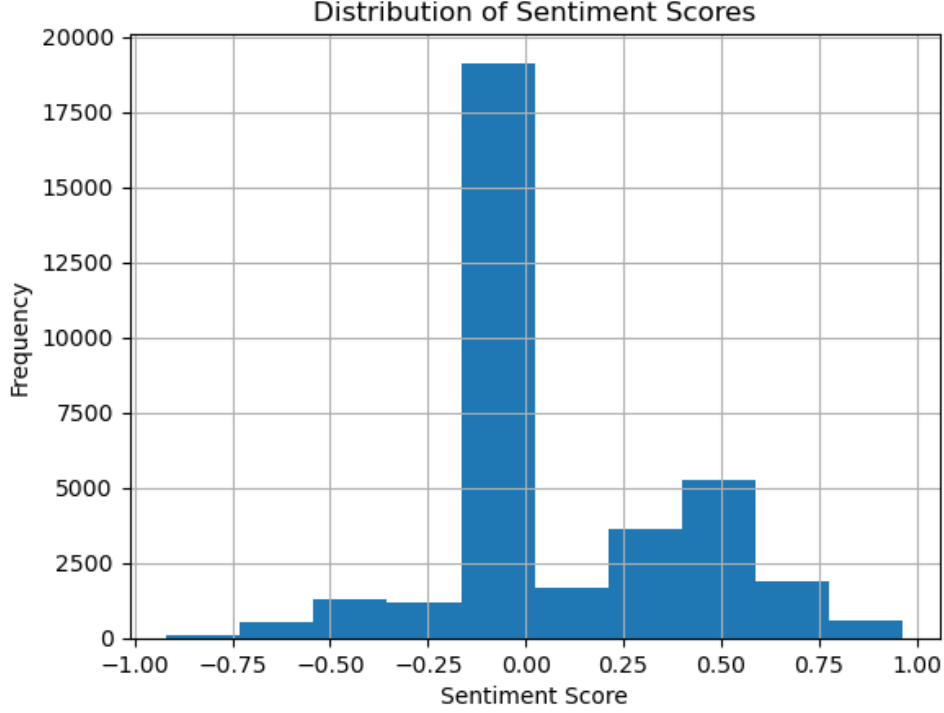
Figure 1: Distribution of Sentiment Scores

This serves as a motivation for adjusting bias weights, as discussed in Section 3.2, which account for market 'hypes' and form the basis for the proposed hype-adjusted probability measure. Bias weights are adjusted to correct for distortions such as excessive coverage of certain stocks and underrepresentation of others. These adjustments ensure a more objective and balanced assessment of market dynamics. Without such corrections, the analysis may disproportionately emphasize certain stocks or news sources, resulting in skewed forecasts or inaccurate conclusions.

3.2 Bias, Memory, and Weights

This section explores the foundational elements - bias, memory, and weights - that underpin the construction of a refined sentiment equation, serving as the basis for the proposed hype-adjusted probability measure. The goal is to systematically address and correct biases arising from media attention imbalances and market weighting discrepancies while incorporating the dynamic influence of past sentiment.

One significant limitation of the prior sentiment equation 3 is that it only counts the number of positive, negative, and neutral news. A news of sentiment score -3 should carry a bigger weight than a -0.1 score.

We propose a modified approach to sentiment scoring. First, an interval of $(-0.05, +0.05)$ is chosen to broaden the range of sentiment scores classified as neutral, setting all scores within this domain to 0. Additionally, sentiment scores of extreme values, such as $+4$, are given higher weights than minimal positive values (e.g., $+0.1$), addressing the limitation in the original model where they had equal weights.

$$Sent_d = \frac{1}{n} \sum_{i=1}^{n} SentScore_{i,d} \tag{5}$$

Notations

1. n: The total number of tickers in the portfolio.

2. $Sent_d$: The compound daily sentiment score of an underlying asset portfolio, calculated without considering historical data.

3. $SentAll_d$: The compound daily sentiment score incorporating the cumulative influence of historical data.

4. $SentScore_{i,d}$: The average sentiment score for ticker i on date d, calculated from a total of k news articles' sentiments $S_{i,d,j}$.

$$SentScore_{i,d} = \frac{1}{k} \sum_{j=1}^{k} S_{i,d,j} \tag{6}$$

3.2.1 News Bias

News bias refers to the disparity in media coverage that individual stocks or sectors receive, which may not align with their actual market significance. For instance, major companies like Nvidia tend to dominate media attention, potentially distorting the sentiment analysis for the entire sector.

The ticker news count weight represents the proportion of total news articles attributed to a specific ticker compared to all tickers within the sector. This metric quantifies the relative attention a ticker receives from the media and audience in comparison to other tickers in the same domain or sector. Formally, for ticker i:

$$\text{Ticker news count weight}_i = \frac{\text{News count}_i}{\sum_{j=1}^{n} \text{News count}_j}, \tag{7}$$

where n is the total number of tickers in the sector. This weight reflects how much media coverage a particular ticker receives relative to others.

The new bias arises from the gap between the Ticker news count weight and the actual market weight of a ticker. Mathematically:

$$\text{Bias}_i = \text{Ticker news count weight}_i - \text{Market weight}_i, \tag{8}$$

where:

- A positive bias ($\text{Bias}_i > 0$) indicates over-representation in the news.

- A negative bias ($\text{Bias}_i < 0$) suggests under-representation in the news.

Adjusting for this bias is essential to ensure that the sentiment analysis accurately reflects market dynamics rather than being skewed by excessive coverage or neglect of specific tickers. Without such adjustments:

- Over-represented tickers (e.g., Nvidia) may disproportionately influence sector sentiment, leading to inflated predictions.

- Under-represented tickers may not properly contribute to the overall sentiment.

By integrating both Ticker news count weight and Market weight into the sentiment equation, we aim to balance these disparities and improve the accuracy of market sentiment models.

We examine news bias within the iShares Semiconductor ETF (SOXX) of the U.S. stock market by analyzing how individual asset components contribute to the overall index. Several major companies, such as Nvidia, dominate the sector, with Nvidia representing over 8% of the market. However, over the previous 15 months (up to mid-July 2024), Nvidia had received disproportionately high news coverage of 24.52% compared to other companies in the sector.

Considering 30 companies from the SOXX, we assigned weights to each ticker based on their market share. For example, Nvidia (NVDA), valued at $1,120,866.62 million and accounting for 8.64% of the sector, has a component weight of 8.64%. Please refer to Appendix A for more details.
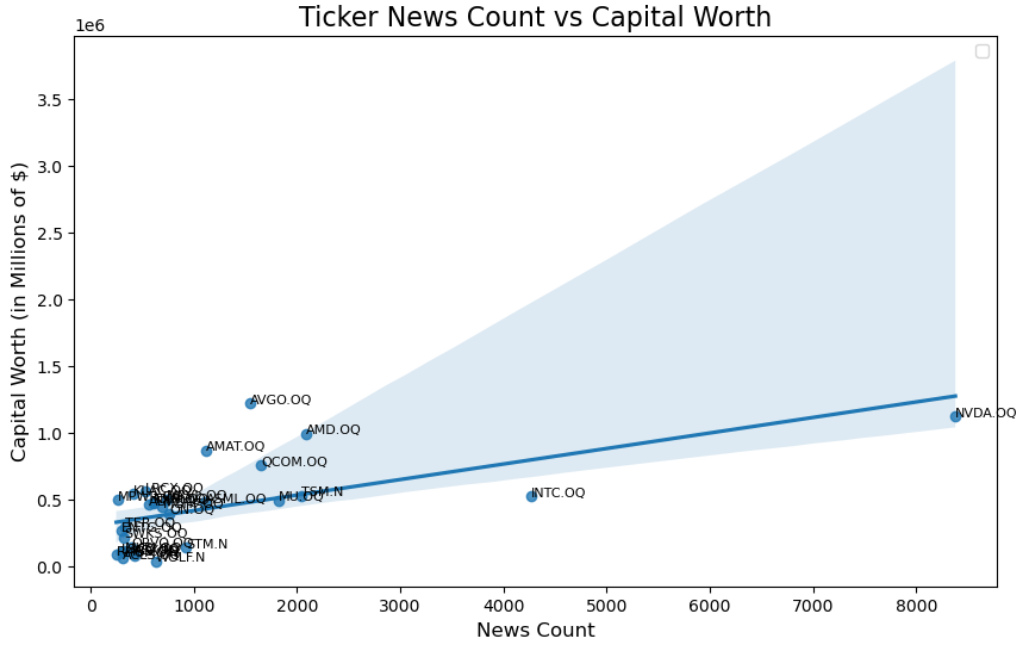

Figure 2: Actual Linear Relationship of Ticker news count weight vs Capital worth weight

A zoom on the left bottom side of the plot is provided by filtering out NVDA.QQ and INTC.QQ from the ticker list.
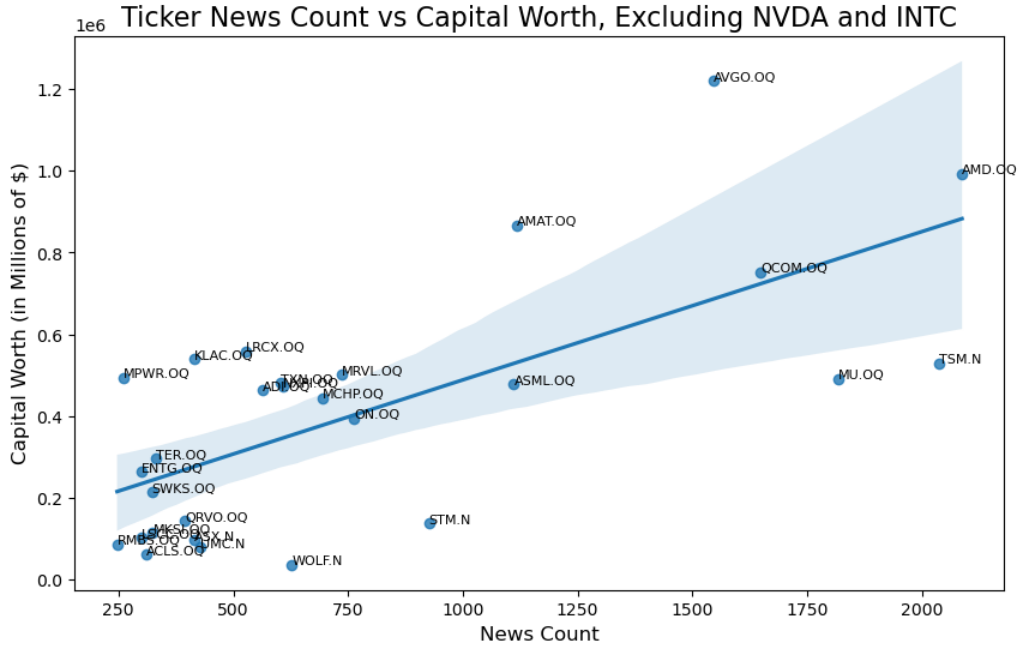

Figure 3: Actual Linear Relationship of Ticker news count weight vs Capital worth weight, Excluding NVDA and INTC

We wish to provide a sophisticated method to change the relationship from the first plot in Figure 2 to the second in Figure 4.
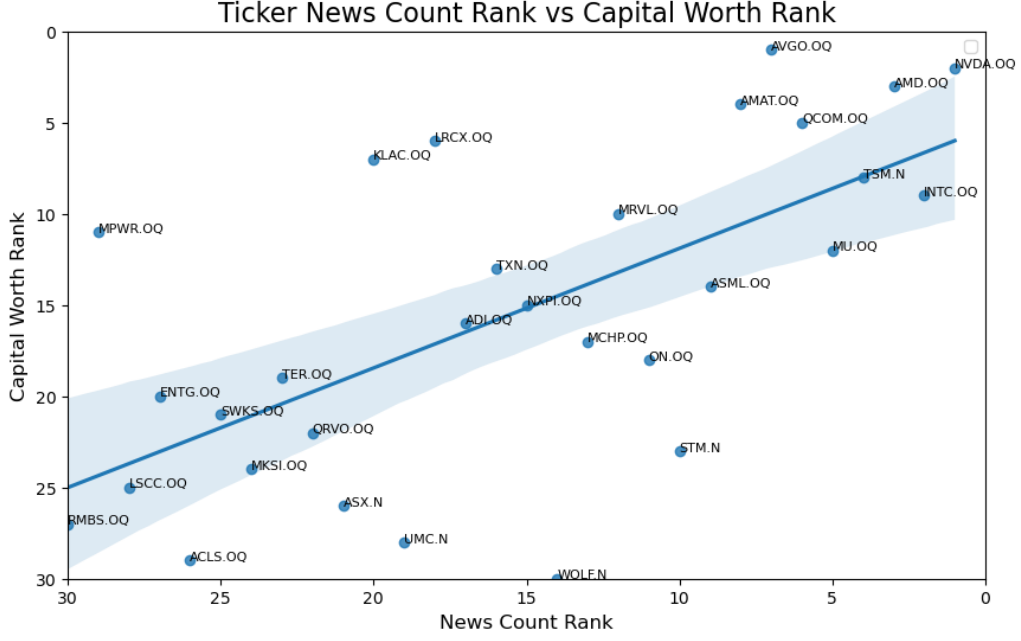
Figure 4: Ideal Linear Relationship of Ticker News Count Rank vs Capital Worth Rank

3.2.2 Memory and Weight

Quantifying the bias of news sources and optimizing compound sentiment scores are key to improving sentiment-based forecasting models. We propose new parameters for calculating the daily compound sentiment score. Rather than using daily averages in equation 6 as a benchmark for $Sent_d$ (sentiment score on the date or time d), we assign weights to each component (ticker) based on its sector weight:

$$Sent_d = \sum_{i=1}^{n} \omega_i \, SentScore_{i,d} \tag{9}$$

Here:

- d: represents the current day.

- n: denotes the total number of tickers studied.

We introduce a novel approach to account for time-memory effects in market reactions. Historical news sentiment impacts future market performance, with older news exerting diminishing influence. For instance, the collective sentiment on a Monday may be shaped not only by weekend news but also by earlier events. We propose weighting algorithms to capture these decay and lagging effects, ensuring older events carry less influence but some persistence.

$$Sent_d = \omega_{\text{today weight}} \sum_{i=1}^{n} \omega_{\text{component weight}} SentScore_{i,d} \tag{10}$$
$$+ \omega_{\text{past weight}} Sent_{d-1}$$

This equation includes:

- Present sentiment: A weighted sum of sentiment scores adjusted for news bias and component weights $(\omega_{\text{component weight}})$.

- Past sentiment: The sentiment from the previous day $(Sent_{d-1})$ is included, weighted by its relevance over time.

By incorporating these weights, the model ensures that the sentiment score reflects both the immediate impact of current news and the lingering effects of past sentiment, thereby improving the accuracy of market forecasts.

The primary goal is to determine the extent to which news sources exaggerate or understate events and how this impacts market sentiment. By incorporating bias weights, memory effects, and component-specific adjustments, the model aims to better reflect true market sentiment, minimizing the effects of biased reporting. Future improvements could involve more advanced bias detection algorithms and real-time integration of data from platforms like Eikon.

Following equation 10 and using the semiconductor sector as an example, we analyze how individual asset components contribute to the overall index. Considering 30 companies from the SOXX, we assign weights to each ticker based on their market share. For example, Nvidia (NVDA), valued at \$1,120,866.62 million and accounting for 8.64% of the sector, has a component weight of 8.64%.

$$\sum_{n=0}^{29} \omega_n = 1 \tag{11}$$

We propose that shifts in sentiment direction affect the relevance of historical data. This is incorporated into the construction of indicator functions in the new sentiment score equation, adjusting the weight of past data based on directional changes.

3.2.3 The 3 Criteria

We examine the impact of three types of weights on market sentiment analysis:

1. Market Component Weights vs News Coverage:

Certain stocks receive disproportionate media coverage compared to their actual weight in the market or sector. For instance, Nvidia may garner excessive news coverage relative to its actual weighting in the semiconductor sector.

To address this imbalance, we propose a solution that adjusts the contribution of individual tickers' sentiment to the overall sentiment of the industry or sector. This approach aims to mitigate the disproportionate influence of certain stocks, progressing from the current scenario (top figure) to a more balanced perspective (bottom figure).

Currently, the approach is to train the algorithm to learn the best weight for each individual component

2. News Source Bias Weights:

We recognize that news reports are rarely perfectly objective. Different news sources have biases that can affect sentiment. For example, CNN may exhibit a liberal bias and present a relatively positive outlook on green energy, whereas FOX may be more critical of gun control policies. Similarly, the BBC may adopt a more critical stance on the Chinese market compared to state-owned Chinese media, with biases shaped by the source's audience and language.

To correct for these biases, we propose assigning weights to different news sources based on the degree of bias and applying these weights when analyzing sentiment data.

To quantify the news bias, we plan to use data from different media sources and predict the results separately. We assume that a more accurate forecast comes from a more objective data source. And we will apply the learned degree of bias to assign new weights to compute the adjusted sentiments.

3. Weights of Past News Data (Memory):

We have already demonstrated that adjusting sentiment using indicator functions, coupled with ML techniques to optimize parameters, can enhance the accuracy of market forecasts.

Building on this, we aim to derive a new score equation that captures how the weighting of past news data (memory) influences current prices. This weighting function will help us better understand how historical sentiment data affects current market dynamics.

In summary, we focus on evaluating and optimizing the 3 criteria to enhance the sentiment score equation and NLP approach. These enhancements ensure a balanced and accurate representation of market dynamics by correcting for media attention imbalances, accounting for source biases, and incorporating memory effects. This foundational framework not only refines sentiment modeling but also provides the necessary groundwork for the development of an adjusted sentiment score equation. In the next section, we extend these concepts to introduce a more robust equation that integrates component weighting and historical memory effects for greater predictive accuracy.

3.3 Adjusted Sentiment Score Equation

This section presents an enhanced sentiment score calculated based on equation 10 that incorporates the weighting of individual components within a portfolio and accounts for historical memory effects from past news data.

We incorporate only the most recent significant sentiment shift event. If multiple shifts have occurred but the threshold for disregarding prior data has not been met, only the most recent shift is considered, and sentiments from earlier dates are neglected.

$$
\begin{aligned}
SentAll_d = & \sum_{i=1}^{k} \mathbb{1}_{Sent_d, Sent_{d-1}} \quad \omega_{\text{weight}_{i,d}} SentScore_{i,d} \\
= & \omega_{\text{today weight}} Sent_d \\
& - \mathbb{1}_{Sent_d < 0 \ \& \ Sent_{d-1} > 0} \quad \omega_{\text{past weight 1}} SentAll_{d-1} \\
& + \mathbb{1}_{Sent_d > 0 \ \& \ Sent_{d-1} < 0} \quad \omega_{\text{past weight 2}} SentAll_{d-1} \\
& + \mathbb{1}_{Sent_d \times Sent_{d-1} \geq 0} \quad \omega_{\text{past weight 3}} SentAll_{d-1}
\end{aligned}
\tag{12}
$$

We consider the influence of historical sentiment. The $SentAll_d$ variable represents the compound daily sentiment score, taking into account the cumulative influence of historical data. Each term in this equation has the following interpretation:

We categorize sentiment changes into three cases to assess the role of historical data in the overall sentiment function, based on indicator functions:

1. $\mathbb{1}_{Sent_d < 0 \ \& \ Sent_{d-1} > 0}$: when the daily sentiment shifts from positive to negative.

2. $\mathbb{1}_{Sent_d > 0 \ \& \ Sent_{d-1} < 0}$: when the daily sentiment shifts from negative to positive.

3. $\mathbb{1}_{Sent_d \times Sent_{d-1} \geq 0}$: when the daily sentiment directions stay unchanged

When the indicator $\mathbb{1}_{Sent_d < 0 \ \& \ Sent_{d-1} > 0}$ is met, higher volatility is likely to occur as compared to the condition $\mathbb{1}_{Sent_d > 0 \ \& \ Sent_{d-1} < 0}$.

The real-world implication leads us to hypothesize that the more positive the overall sentiment is (the greater $SentAll_{d-1}$ is), the more impacts it will cause from negative news occurrences to traders' panic sales and the drop in market prices (the smaller $SentAll_d$ becomes, and therefore introducing the negative sign before $\mathbb{1}_{Sent_d < 0 \ \& \ Sent_{d-1} > 0}$.

We hypothesize that when sentiment shifts from positive to negative, the impact becomes more significant as the magnitude of the change increases. This would directly correlate with the weight we aim to introduce here.

Weights below 0.005 are disregarded. For instance, after multiple recursive applications, if $\omega_{\text{time weight}, d-10} = 0.003 < 0.005$, data for dates on or before $d - 10$ are excluded from consideration.

In the following section, we apply the enhanced sentiment score equation within an updated NLP framework, testing its predictive power using various algorithms and assessing its impact on market forecast accuracy.

3.4 New NLP Approach and Forecast Results

This section summarizes the ML methodologies, procedures, and forecast results used in the new NLP approach, which is based on the enhanced sentiment score equation.

3.4.1 Methodology

The primary method introduced in Deveikyte et al relied on Latent Dirichlet Allocation [6]. In contrast, we use in this study unsupervised ML models as benchmarks to compare the performance of enhanced computational and optimization techniques.

Our objectives are twofold: first, improve the forecasting accuracy of the original model, and second, demonstrate the benefits of incorporating arithmetic computations and optimizations beyond standard LDA.
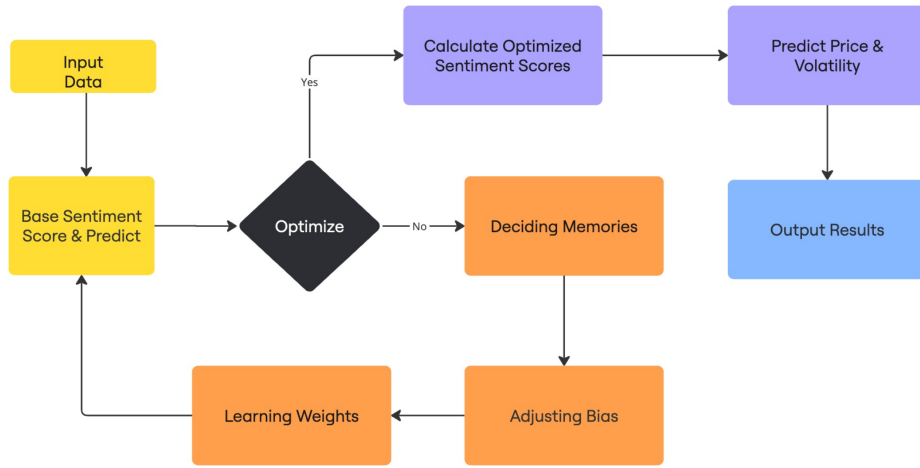
Figure 5: Algorithm Flow Chart

The optimization algorithms used include models such as LDA, Logistic Regression, Ordinary Least Squares (OLS) regression. The primary focus of this paper is not to fine-tune ML models but to demonstrate the improvements gained from the new sentiment score data and procedure.

3.4.2 Machine Learning Procedures

We employ several ML approaches to predict stock market movements based on sentiment data. The sentiment data is first processed by converting the date index into a standardized date format. This data is then merged with stock market data, including daily market returns and volatility metrics.

The methodology involves iterating through various sentiment indicators and performing data splits into training, validation, and test sets. For each indicator, we run the model up to 1,000 times with different random states to ensure robust results.

Ordinary Least Squares (OLS) regression is used to quantify the relationship between sentiment scores (independent variable X) and daily market return or volatility (dependent variable Y). The regression provides an R-squared value to measure the proportion of variance in Y explained by X. This allows us to evaluate the predictive power of sentiment scores in explaining market dynamics.

Logistic regression is employed to predict the direction of daily market returns (dependent variable Y) based on the compound sentiment score (independent variable X). Here, Y is a binary variable indicating positive or negative market movements. The accuracy of the logistic regression model serves as a key metric for assessing its performance.

The optimal models are selected based on the highest validation accuracy (for logistic regression) and R-squared values (for OLS regression) across iterations. These models are subsequently evaluated on the test dataset, with metrics including accuracy, precision, recall, and confusion matrices. The results highlight the most effective sentiment indicators for predicting market movements, offering insights into the relationship between sentiment data and market behavior.

Table 4 shows a sample of modified sentiment scores based on different parameters in equation 12.

Table 4: Sample Sentiment Scores with Selected Parameters

Date	Base Sentiment	Parameters A	Parameters B	Parameters C	...
2023-04-27	0.043192	0.043192	0.043192	0.043192	...
2023-04-28	0.186383	0.186383	0.207979	0.229574	...
2023-04-29	0.040962	0.040962	0.144951	0.270536	...
2023-04-30	-0.036497	-0.036497	-0.036497	-0.036497	...
2023-05-01	0.078847	0.078847	0.078847	0.078847	...

Each column represents the adjusted sentiment scores based on different sets of parameters. The model results are analyzed in Section 5.

4   A Hype-Adjusted Probability Measure

In the context of financial markets, hype refers to the amplification of attention or sentiment around a particular stock, sector, or market event that exceeds its fundamental or intrinsic importance. Hype is often fueled by disproportionate media coverage, speculative behavior, or investor overreaction to news.

A hype scenario is identified when there is a measurable and disproportionate increase in examples like:

1. Media Coverage: A significant spike in the volume of news articles, social media mentions, or other sources of information about a specific stock or sector compared to its baseline or relative importance.

2. Market Over Reactions: Corresponding anomalies in price movement and volatility, such as sharp increases or dramatic swings, that deviate from historical patterns.

3. Imbalance in Representation: Evidence of over or under-representation in news coverage compared to the stock's weight in its sector (e.g., market capitalization).

These indicators form the foundation for quantifying and incorporating hype into the proposed probability measure. A real-world case of Nvidia's hype is examined later in this section.

4.1   Intuition of the New Measure

Recall the 3 criteria:

- Market Component Weights vs news Coverage"

- News Source Bias Weights

- Weights of Past News Data (Memory)

These parameters capture the influence of market components, news biases, and historical data retention (memory) on sentiment and subsequently on the market forecast.

Also, recall that we have shown from figure 1 that a slight negative skew of market news sentiments is observed and there exists an imbalance of news report counts vs. component weights of Nvidia among the semiconductor sector tickers.

In addition, the news source bias weights are currently being investigated for further research. For example, in America, major news media have political leanings, CNN is more left-wing and promotes clean energy and green targets, while Republican-favored Fox argues for the opposite tone. We wish to quantify the bias instead of relying solely on polls and surveys.

To illustrate the concept of market 'hype', we examine Nvidia, the most trending AI company as a case study for summer 2024.
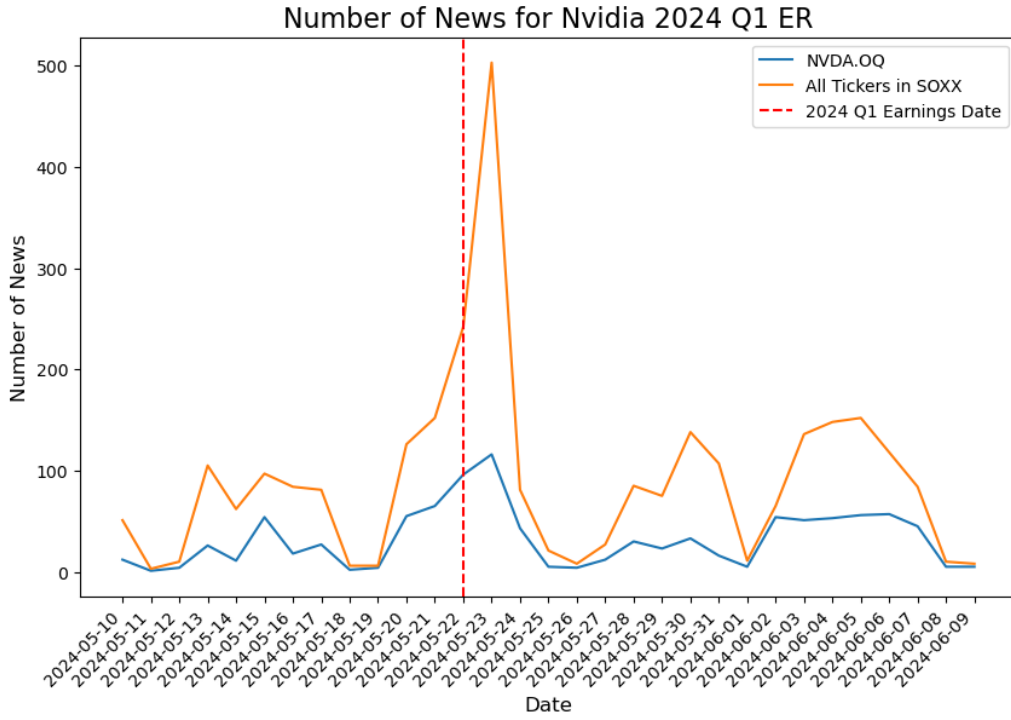
Figure 6: Number of News for Nvidia Q1 Earning Report on May 22, 2024, versus tickers in SOXX

In figure 6, the blue line indicates the news counts on Nvidia from May 10 to June 9, 2024, while the red dashed line indicates May 22, 2024, when Nvidia's Q1 earnings report was released. The orange trajectory describes the total number of news collected for all tickers in the SOXX sector. We notice a rapid increase in market hype centering around the report release time.
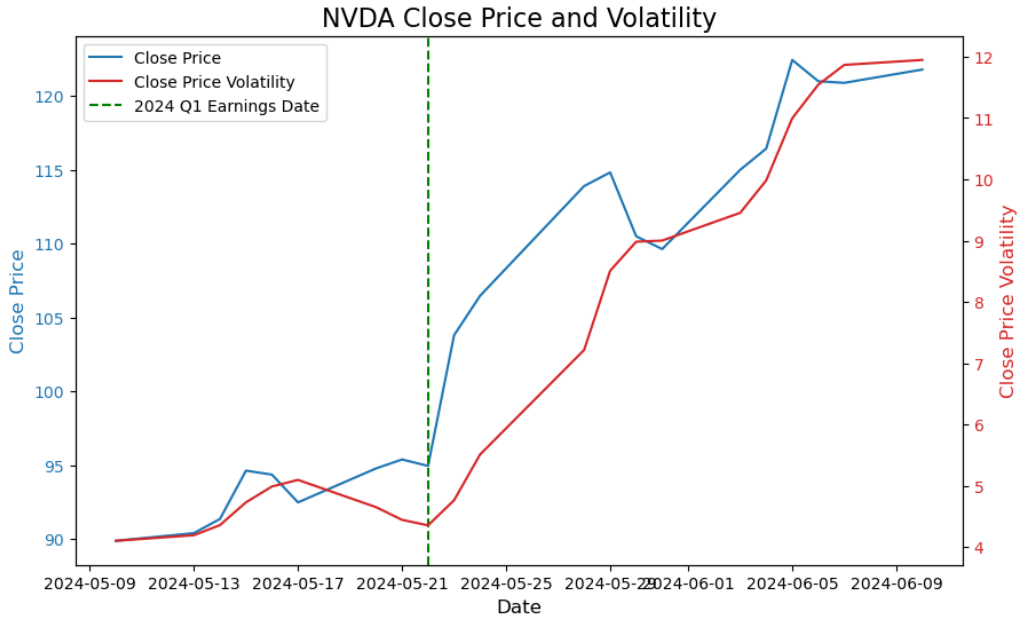

Figure 7: Price and Volatility Trajectories for Nvidia Q1 Earning Report on May 22, 2024

Figure 7 presents the historical movements of Price and Volatility based on the market close data for the same time frame of Figure 6, it is on observation of a dramatic increase in both the close price and the close price volatility.

Notice, while we do not assert a direct positive correlation between news count increases and price or volatility changes, these observations suggest that market hype can significantly

impact these variables, which inspires the proposal of a hype-adjusted probability measure, $\mathbb{P}^a$.

4.2   Construction of A Hype-Adjusted Probability Measure

Based on the previous intuitions, we define a hype-adjusted probability measure, $\mathbb{P}^a$, to account for market sentiment and the existence of news biases.

Hype-adjusted Probability Measure:

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the set of states of nature ω, $\mathcal{F}$ is the filtration of information, and $\mathbb{P}$ the physical probability measure. We define a new probability measure $\mathbb{P}^a$ on $(\Omega, \mathcal{F})$ by assigning new weights to the states of nature, where the effects of news weight and bias from media sources are corrected, reducing the over / under representation in the news of some particular stocks of a sector.

We get the economic inspiration of change of measure from the physical measure $\mathbb{P}$ to a Harrison and Kreps (1979) "risk-neutral" / equivalent-martingale measure; and Geman's Forward measure (1989). Note that in these two famous cases like in our setting below, the change of measure and its weights reflect the economic problem to address.

We suppose that there are 3 states of nature $\omega_1, \omega_2, \omega_3$ in the universe Ω, labeled as Up, Level, and Down.

We know from measure theory that the relationship between a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ and an equivalent probability measure $\mathbb{P}^a$ has the following form:

$$Z = \frac{d\mathbb{P}^a}{d\mathbb{P}}, \tag{13}$$

$$Z(\omega) = \frac{\mathbb{P}^a(\omega)}{\mathbb{P}(\omega)}. \tag{14}$$

where Z is a random variable with

$$\mathbb{E}_{\mathbb{P}}[Z] = 1. \tag{15}$$

In order to define Z, we need to choose $Z(\omega_1)$, $Z(\omega_2)$, and $Z(\omega_3)$.

We consider now two companies, Nvidia and Intel, and their respective stock and sentiment performances:

- Nvidia: Today, the stock increased by 10%, but the sentiment increased by 20%. Nvidia is thus considered "overhyped", we choose:

$$Z(\omega_1) < 1$$

- Intel: The stock decreased by 20%, but the sentiment only decreased by 10%. Intel is thus considered "underhyped", we choose:

$$Z(\omega_3) > 1.$$

1. Up State (w_1):

For Nvidia, which is overhyped, we choose $Z(\omega_{\text{up}}) < 1$, reducing the probability of the Up state. This reflects the fact that the market sentiment is too optimistic relative to the actual stock performance.

$$\mathbb{P}^a(\omega_{\text{up}}) = Z(\omega_{\text{up}}) \cdot \mathbb{P}(\omega_{\text{up}}),$$

such that

$$\mathbb{P}^a(\omega_{\text{up}}) < \mathbb{P}(\omega_{\text{up}}).$$

2. Down State (w_3):

For Intel, which is underhyped, we choose $Z(\omega_{\text{down}}) > 1$, increasing the probability of the Down state to reflect the understated sentiment relative to the actual performance:

$$\mathbb{P}^a(\omega_{\text{down}}) = Z(\omega_{\text{down}}) \cdot \mathbb{P}(\omega_{\text{down}}),$$

such that

$$\mathbb{P}^a(\omega_{\text{down}}) > \mathbb{P}(\omega_{\text{down}}).$$

3. Level State (w_2):

In the "Level" state, sentiment and stock price changes are neutral, and no adjustment is needed. The value we assign to $Z(\omega_{\text{Level}})$ is such that

$$\mathbb{E}_{\mathbb{P}}[Z] = 1.$$

The same construction can be extended to include more stocks.

This adjustment allows us to account for market hype by converting the original news weight of a particular ticker to its correct proportion within the sector. The following section 5 presents the improved NLP forecasting results using a hype-adjusted probability measure.

5 Model Results and Discussion

The results presented in this section highlight the effectiveness of the proposed sentiment-based framework in improving market return and volatility forecasting. By integrating advanced sentiment modeling techniques, we observe substantial accuracy improvements, which pave the way for innovative theoretical contributions and practical applications in financial analysis.

Under a hype-adjusted probability measure of the selected date range, we calculate that the expected average difference between the original volatility and the adjusted one, based on the weights provided in the semiconductor dataset, is approximately -0.0068 (or -0.68%). We note however, that despite this small number, the accuracy of the prediction was improved by 8%.
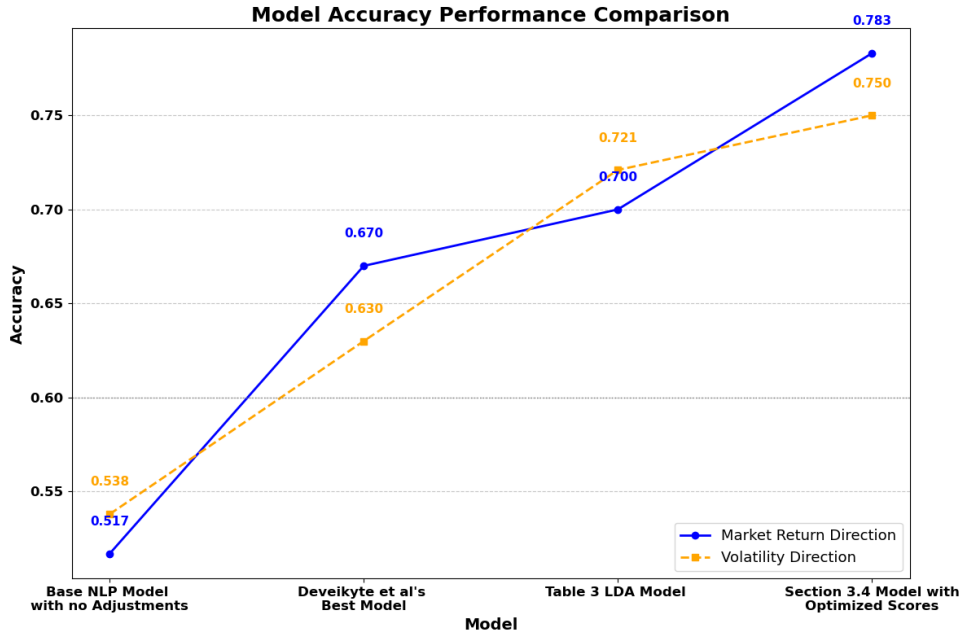

Figure 8: Model Accuracy Comparison for Return and Volatility Forecasting (discussed below)

We choose volatility and market return directions (increase or decrease) as the target of prediction, which aligns with the observation results from sections 3.4 and 4.1.

Figure 8 demonstrates the progression of accuracy across different models. In the prediction of market return direction, validation accuracy improves from 51.7% in the baseline model to 70.0% with a LDA model using adjusted scores, and further to 78.3% with the optimized scores. Similarly, for volatility direction, accuracy increases from 53.8% in the baseline model to 72.1% and 75.0% in the corresponding models. These substantial improvements validate the effectiveness of our methodology.

The observed accuracy increase of +8.3% for market return direction and +2.9% for volatility direction highlights the refined precision achieved through optimized sentiment modeling. This improvement translates into more reliable predictions of market trends, a critical factor for decision-making during high-volatility scenarios and market crises.

Note that we do not claim the uniqueness of the hype-adjusted probability measure $\mathbb{P}^a$ we are proposing, for two sets of reasons:

1. The way we define "hype" can vary based on the approach used to link news to market sentiment. For instance, other researchers may define hype using different NLP techniques, sentiment scoring models, or thresholds. There is no canonical way of defining hype, and our chosen approach is just one of many possibilities to bridge news sentiment with market behavior.

2. Moreover, there could be multiple valid constructions of $\mathbb{P}^a$ that align with the sentiment adjustments.

The flexibility in both defining hype and constructing the adjusted measure ensures the generality of the approach, allowing it to be adapted to various datasets and market contexts.

Regarding our potential future research, one avenue could be to derive and formalize a hype-adjusted volatility and option pricing under a hype-adjusted probability measure. Future work can explore integrating sentiment-guided adversarial learning frameworks, such as Long Short-Term Memory (LSTM) networks and generative adversarial networks (GANs) presented, for instance, by Zhang et al in [1], to enhance the ability of the model to adapt to dynamic market conditions. Further analysis of bias or subjectivity can also be conducted using tools like the Eikon Data API, as discussed earlier in the paper.

An important point to keep in mind is that market participants do not only trade on news. Some trade on the basis of technical analysis, such as moving averages (for commodities but for equities as well); a small category of market participants, called arbitragers, trade only if they have identified a strict arbitrage opportunity; some major hedge funds try to recognize "statistical arbitrage" patterns. Lastly, a category of fundamental analysts uses Capital Structure and Earnings as trading signals. All these strategies interact with each other, leading to market prices whose exact formation is beyond the scope of this paper.

6 Conclusion

In this paper, we present an improved NLP approach for forecasting stock return and volatility based on a hype-adjusted probability measure, $\mathbb{P}^a$.

Besides validating improvements in financial forecasting, these results lay the foundation for broader theoretical applications. A hype-adjusted probability measure is introduced to quantify and integrate market hypes, extending the framework beyond classical sentiment analysis.

Our model evaluates text tokens with greater accuracy, effectively capturing the influence of correlations and weights across subsets. A key advantage is its low sensitivity to the accuracy of the initial token scoring method, making it robust across varied data sources. The enhanced performance of our sentiment model is exhibited in the key result figure which shows the higher accuracy obtained in predicting market responses compared to baseline models.

Lastly, our hype-adjusted probability measure can arguably be valued as a theoretical bridge between the probabilistic finance of Asset Pricing and NLP prediction in Finance, two fields which are likely to intersect more and more in the near future.

References

[1] Yiwei Zhang et al. "Sentiment-Guided Adversarial Learning for Stock Price Prediction". In: *Frontiers in Artificial Intelligence* 7 (2021).

[2] Malcolm Baker and Jeffrey Wurgler. "Investor Sentiment and Stock Returns". In: *Journal of Finance* 61.4 (2006), pp. 1645–1680. DOI: 10.1111/j.1540-6261.2006.00885.x.

[3] Suresh Balakrishnama and Aravind Ganapathiraju. "Linear Discriminant Analysis: A Brief Tutorial". In: *Institute for Signal and Information Processing* 18 (1998), pp. 1–8. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=aJjWmjUAAAAJ&citation_for_view=aJjWmjUAAAAJ:GnPB-g6toBAC.

[4] Yoshua Bengio et al. "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3 (2003), pp. 1137–1155.

[5] Dror Cohen et al. "Masking important information to assess the robustness of a multimodal classifier for emotion recognition". In: *Frontiers in Artificial Intelligence* 6 (2023). ISSN: 2624-8212. DOI: 10.3389/frai.2023.1091443. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1091443.

[6] J Deveikyte et al. "A sentiment analysis approach to the prediction of market volatility". In: *Front. Artif. Intell.* 5 (2022), p. 836809. DOI: 10.3389/frai.2022.836809.

[7] Peter Gabrovsek et al. "Twitter sentiment around the earnings announcement events". In: *PLoS ONE* 12 (2016), e0173151. DOI: 10.1371/journal.pone.0173151.

[8] Hélyette Geman. "The Importance of the Forward Measure for the Pricing of Interest Rate Derivatives". ESSEC Working Paper. 1989.

[9] Paul Glasserman and Harry Mamaysky. "Does Unusual News Forecast Market Stress?" In: *Management Science* 63.10 (2017), pp. 3397–3414. DOI: 10.1287/mnsc.2016.2513.

[10] London Stock Exchange Group. *LSEG Data and Analytics*. Collected from LSEG on July 17, 2024. July 2024. URL: https://www.lseg.com/.

[11] J. Michael Harrison and David M. Kreps. "Martingales and arbitrage in multiperiod securities markets". In: *Journal of Economic Theory* 20.3 (June 1979), pp. 381–408. URL: https://ideas.repec.org/a/eee/jetheo/v20y1979i3p381-408.html.

[12] David Hirshleifer. "The Role of Sentiment in Financial Markets: A Survey of the Theoretical and Empirical Literature". In: *Annual Review of Financial Economics* 1.1 (2001), pp. 132–159. DOI: 10.1146/annurev.financial.1.080801.100352.

[13] Clayton J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (2014), pp. 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[14] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd. Prentice Hall, 2000.

[15] Bing Liu. *Sentiment Analysis and Opinion Mining*. Vol. 5. Synthesis Lectures on Human Language Technologies 1. Morgan & Claypool, 2012. DOI: 10.2200/S00416ED1V01Y201204HLT016. URL: https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

[16] Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. "Measuring News Sentiment". In: *Journal of Econometrics* 228.2 (2020). Federal Reserve Bank of San Francisco, United States of America; Motive Software, United States of America, pp. 221–243. DOI: 10.1016/j.jeconom.2020.07.005.

[17]   Ning Zhong. "Sentiment, Trading Behavior, and Investors' Decision-Making". In: *Review of Financial Studies* 31.1 (2018), pp. 123–151. DOI: 10.1093/rfs/hhx103.

A   Appendix A: Ticker News Weight Table

For the table below, *Close Price* refers to the price per share at market close in U.S. dollars; *Capital* is reported in millions of dollars; *Capital Weight %* represents the company's market capitalization as a percentage of the entire sector's capitalization; and *News Weight %* indicates the proportion of news coverage the company receives relative to the whole sector.

Note, in the first row, the remarkable difference between columns 3 and 4.

Ticker Name	Close Price	Capital	Capital Weight %	News Weight %
NVDA.OQ	131.38	1120867.0	8.66	24.52
INTC.OQ	34.59	524362.2	4.05	12.49
AMD.OQ	177.1	992494.8	7.67	6.11
TSM.N	184.52	528702.9	4.08	5.96
MU.OQ	131.14	490637.9	3.79	5.32
QCOM.OQ	207.12	752063.7	5.81	4.82
AVGO.OQ	1733.31	1220293.0	9.43	4.52
AMAT.OQ	251.47	866274.3	6.69	3.27
ASML.OQ	1059.97	477667.0	3.69	3.25
STM.N	41.51	137307.8	1.06	2.71
ON.OQ	73.48	394591.1	3.05	2.23
MRVL.OQ	73.84	501852.5	3.88	2.15
MCHP.OQ	92.34	444145.9	3.43	2.03
WOLF.N	23.05	35874.98	0.28	1.83
NXPI.OQ	274.91	472496.2	3.65	1.78
TXN.OQ	200.16	480016.7	3.71	1.77
ADI.OQ	232.01	462719.8	3.57	1.65
LRCX.OQ	1112.55	558002.5	4.31	1.54
UMC.N	8.59	78866.05	0.61	1.25
KLAC.OQ	874.9	538704.0	4.16	1.22
ASX.N	11.96	96245.99	0.74	1.21
QRVO.OQ	119.69	142502.9	1.10	1.15
TER.OQ	153.48	295777.2	2.28	0.97
MKSI.OQ	135.06	113224.3	0.87	0.95
SWKS.OQ	106.41	213374.7	1.65	0.94
ACLS.OQ	151.06	60574.45	0.47	0.91
ENTG.OQ	140.23	263961.8	2.04	0.88
LSCC.OQ	59.75	101910.3	0.79	0.87
MPWR.OQ	846.2	493544.3	3.81	0.76
RMBS.OQ	63.88	85799.32	0.66	0.72

Note, this table is collected from LSEG, on July 17, 2024[10]. This ticker weight table is adjusted by removing 2 values: CME E-MINI S&P500-TECHNOLOGY SECTOR INDEX FUTURE SEP 2024, 0.15247% and CME INDEX and OPTIONS MARKET E-MINI RUSSELL 2000 INDEX FUTURE SEP 2024, 0.15247%.