Markets Have Memory – But What About Mood?

Forecasting Volatility with Sentiment-Augmented HAR Models

Max Moore* Clara Nielsen†

MSc in Finance
Stockholm School of Economics
May 2025

Abstract

This thesis examines whether sentiment-based indicators improve forecasts of daily volatility for the S&P 500. Using a large panel of investor, public, and institutional sentiment measures, we extend the heterogeneous autoregressive (HAR) model to incorporate sentiment information. Empirical results show that sentiment variables significantly enhance both in-sample and out-of-sample predictive performance, with investor sentiment proxies - particularly option skew and bullish positioning, emerging as the most reliable predictors. Threshold analysis reveals that the predictive impact of sentiment is regime-dependent, strengthening during periods of elevated market volatility. These findings contribute to the literature by demonstrating the incremental value of sentiment in volatility forecasting and highlighting the conditions under which it is most effective.

Keywords: Stock Market Volatility, Sentiment Analysis, HAR Models, Volatility Forecasting, Behavioral Finance, Realised Variance, LASSO, Principal Component Analysis.

Supervisor: Tobias Sichert, Assistant Professor, Department of Finance, Stockholm School of Economics.

Acknowledgements: We would like to thank our supervisor Tobias Sichert for his support and guidance throughout the process of writing this thesis.

* 42653@student.hhs.se

† 24815@student.hhs.se

Contents

1 Introduction and Contribution

Forecasting stock market volatility remains a key priority for practitioners and regulators, where reliable forecasts enable better-informed investment decisions, effective hedging, and more robust policy design. While traditional forecasting methods rely on historical returns and statistical time-series models, behavioural finance suggests that investor sentiment - economic participants' collective mood or expectations - could enhance predictive accuracy (Baker and Wurgler, 2007; Yu and Yuan, 2011). In fact, market volatility appears closely tied to psychological forces such as optimism, fear, and uncertainty (Shiller, 2003). Sentiment indicators such as option-implied skew, survey responses, and policy-uncertainty indices capture these behavioural dimensions that conventional models may miss (Lee et al, 2002; Caldara and Iacoviello, 2018).

However, empirical research has yielded mixed conclusions regarding the incremental forecasting power of sentiment-based predictors relative to conventional macroeconomic or market variables. Studies utilising sentiment proxies find them predictive of equity market volatility but differ in robustness, predictive horizons, and sample periods (Tetlock, 2007; Verma and Verma, 2007; Li et al, 2022). Moreover, sentiment is multifaceted and originates from distinct groups within the financial environment - individual investors, broader consumer bases, and institutional actors. These groups may respond differently to market developments and possess divergent informational advantages, influencing volatility at varying intensities and frequencies (Ben-David et al, 2018; Gupta et al, 2023).

This study builds upon existing volatility-forecasting literature in three ways. First, it benchmarks sentiment-augmented specifications against conventional models within an empirical framework, providing direct evidence of sentiment's incremental predictive value. Second, it classifies and compares sentiment measures across investor, public, and institutional domains, identifying the indicators that deliver the greatest practical forecasting gains. Third, it documents the efficacy of sentiment as regime-dependent, clarifying how behavioural signals interact with market conditions and thus extending theoretical discussions within behavioural finance. To our knowledge, this paper is the first to both systematically analyse a broad range of sentiment variables, compare clustered groups, and thoroughly identify key drivers of difference in behaviours across said groups.

Explicitly, this thesis addresses two interconnected research hypotheses:

H1. Sentiment-augmented HAR models outperform traditional forecast models of realised equity market variance, both in- and out-of-sample.

H2. The predictive power of sentiment for volatility will vary by sentiment type (investor, public, institutional) and regime.

This study leverages an extensive dataset spanning multiple sentiment categories and applies statistical forecasting frameworks, including the Heterogeneous Autoregressive (HAR) model augmented with sentiment indicators, regularisation techniques, and regime analysis, to systematically quantify sentiment's predictive value for daily volatility forecasts. By employing forecasting evaluation methods and robustness checks, this thesis assesses whether sentiment information enhances forecast accuracy and discerns the differential impacts of investor, public, and institutional sentiment types on equity market volatility.

The thesis is arranged as follows: Section 2 surveys existing literature, outlining the theoretical and empirical links between sentiment and volatility. Section 3 describes the data and the pre-estimation diagnostics. Section 4 outlines methodological frameworks, emphasising HAR-type models, variable-selection procedures, and forecast-evaluation metrics. Section 5 reports empirical findings, covering both in- and out-of-sample evaluations. Section 6 presents robustness tests and diagnostic checks to assess the reliability of the results. Section 7 analyses the evidence within broader theoretical and practical contexts and discusses various implications. Section 8 concludes by summarising the principal insights and suggesting directions for future research.

2 Literature

The literature review covers overarching theory and literature connected to the relationship between sentiment and stock market volatility, overarching literature regarding forecasting frameworks, as well as gaps in the literature. More in-depth literature regarding the sentiment variables and their connection to volatility is covered in Section 3.

2.1 Theoretical Background

For decades, traditional financial theories have relied on the Efficient Market Hypothesis (EMH), which states that asset prices fully reflect all available information - meaning variables like sentiment should not matter. In this view, irrational behaviour is quickly corrected by rational traders who exploit mispricing and push prices back to equilibrium. However, more recent work, especially from behavioural finance, has started to question this assumption. These models suggest that investor psychology sometimes overwhelms rational forces, limiting arbitrage and allowing price swings and volatility to persist.

One of the most influential behavioural finance models addressing this is the Noise Trader framework developed by De Long et al (1990). The model suggests that the so-called irrational "noise traders" have divergent beliefs, or sentiments, that are unpredictable and influence the pricing of assets. Due to their innate unpredictability, "noise trader risk", or risk of further mispricing, can occur, which limits the possibility for these rational arbitrageurs to fully bet against the mispriced assets. Although the model primarily addresses mispricing, it has implications for volatility as well: having unpredictable investors with sentiment-driven trading adds an additional risk that rational investors must account for when trading, leading to excess volatility. In other words, when investors become excessively optimistic or pessimistic, the consequences in the stock market are more extreme than what the fundamental theories would suggest. While rational investors may eventually bet against the mispricing, the risk that sentiment may intensify in the short term allows such mispricing, and hence elevated volatility, to persist. This suggests that sentiment is a key driver of time-varying risk in equity markets, extending beyond the scope of traditional volatility models. In practice, this has been observed by, for instance, Shiller (2003), who concluded that the EMH fails to account for the increased swings observed during bubbles and crashes, motivating the rise of behavioural explanations.

Furthermore, sentiment can also be understood through the lens of risk aversion and uncertainty. When there is greater uncertainty, investors demand a higher risk premium, which can be linked to volatility through the notion that markets reprice risk. In more stable periods, by contrast, perceived risk tends to be lower, resulting in investors potentially demanding a lower risk premia, reducing volatility. Measures of market mood may therefore contain predictive information about future stock market volatility.

2.2 Empirical Evidence

2.2.1 Sentiment and Volatility

Over the past two decades, the connection between investor sentiment and stock market volatility has attracted increasing attention. Brown (1999) and later Lee et al (2002) were among the first to show that sentiment may influence volatility beyond what is explained by fundamentals. Lee et al, in particular, found a significant relationship between sentiment and conditional variance where periods of bullish sentiment were linked to lower volatility, while bearish sentiment had the opposite effect. Interestingly, this relationship tends to be more pronounced when sentiment is negative. Verma and Verma (2007) supported this, finding that negative news often triggers stronger market reactions and contributes to increased volatility.

Since the early 2000s, studies have continued to explore the effect of sentiment on volatility. Baker and Wurgler (2006; 2007) demonstrated that sentiment plays a role in driving market outcomes through its impact on stock market returns, although evidence is inconclusive as to what extent sentiment anticipates return (Brown and Cliff, 2004). Nonetheless, these studies laid the groundwork for volatility-focused research, as the connection between sentiment and mispricing in returns extends to variance. This was later demonstrated by Yu and Yuan (2011), who found that sentiment influences the mean-variance relation in the stock market. Following this, several studies have also examined the predictive value of sentiment variables in formal volatility models by incorporating sentiment proxies into forecasting frameworks and testing their incremental value. Lee et al (2002), Yu and Yuan (2011), and Bai et al (2024), all conclude that the inclusion of sentiment variables improves the explanatory power of conditional volatility models in equity markets. These studies typically use sentiment indices such as investor surveys, which were statistically significant coefficients in GARCH-type models for volatility, beyond what solely past returns and volatility could predict. This confirms that sentiment is not merely reactive to volatility but may also anticipate it.

A further comparison between market sentiment and macroeconomic variables was conducted by Lindblad (2017), using a GARCH-MIDAS framework in which sentiment variables were incorporated in the low-frequency component of volatility. When tested against the macroeconomic variables, the results showed that once sentiment was included, most macroeconomic variables no longer improved volatility forecasts. This implies that sentiment captures some of the same information reflected in the macroeconomic variables. Sentiment variables have also been shown to outperform macroeconomic variables in explaining volatility movements. Li et al (2022) examined a broad set of potential predictors for S&P 500 realised variance, including economic policy uncertainty indices, market sentiment measures, and various macroeconomic indices. The results demonstrate that market-based sentiment indicators such as VIX, as well as the policy uncertainty indices, outperformed standard macroeconomic variables in their predictive power for stock market sentiment. In this study, VIX exhibited the highest predictive performance, demonstrating that sentiment variables may better capture investors' perceptions of risk and forward-looking expectations, whereas macroeconomic variables tend to reflect realised economic conditions that may lag behind market developments.

As the field has developed, several new sentiment indicators have been identified as having predictive power over stock market prices and volatility. Tetlock (2007) found that high pessimism in *The Wall Street Journal* news articles could predict negative changes in stock market prices. The analysis of media sentiment has since led to the creation of several new sentiment indices with demonstrated predictive power over stock market volatility. One notable example is the Geopolitical Risk (GPR) index, which analyses news articles related to geopolitical uncertainties (Caldara and Iacoviello, 2018).

2.2.2 Forecasting Frameworks for Volatility Models

The aforementioned research has established that sentiment can serve as a relevant predictor for stock market volatility. Early studies, including Paye (2012), primarily relied on simpler predictive regression models when forecasting volatility. More recently, however, structured frameworks have emerged that capture sentiment in a more sophisticated manner. One of the more commonly used models is the Heterogeneous Autoregressive (HAR) model, introduced by Corsi (2009), which incorporates lagged realised variance across multiple time horizons to account for the persistence and clustering of volatility. This model has been extended to include exogenous variables, resulting in a HAR-X variant which enables the evaluation of external predictors such as sentiment variables.

One challenge in introducing multiple exogenous variables into a regression model is the risk of overfitting and multicollinearity, particularly when variables are highly correlated, as could be the case with sentiment variables. To address this, researchers have developed dimensionality-reduction techniques to improve model stability and interpretability. For instance, Li et al (2024) apply LASSO (Least Absolute Shrinkage and Selection Operator), first introduced by Tibshirani (1996), and Audrino et al (2020) apply adaptive LASSO (with variable-specific penalties), developed by Zou (2006). This approach enables variable selection in high-dimensional settings, both improving interpretability and out-of-sample forecast accuracy. Similarly, Asgharian et al (2013) apply principal component analysis (PCA) to address multicollinearity by extracting latent factors from a broad set of predictors.

In addition to solving multicollinearity, the HAR model has also been extended to address the structure of volatility itself. Andersen et al (2007) introduced the HAR-CJ model, which decomposes volatility into two components: continuous and jump. This allows the model to better capture volatility behaviour in turbulent periods where markets are often driven by abrupt movements. The HAR-CJ model has been shown to improve the prediction of volatility shocks that standard models may overlook. Collectively, these models represent the foundation of a growing body of literature supporting the use of the mentioned empirical models in sentiment-augmented forecasting of stock market volatility.

2.2.3 Gaps in Research

While prior literature examined the relationship between sentiment variables and the predictability of market volatility, there are still some areas that remain unexplored.

Comparative analysis of diverse sentiment measures: Most existing literature tends to focus more on a smaller number of sentiment proxies at a time, limiting the opportunity for systematic comparison. Since sentiment is a concept with no universally accepted proxy, a wide range of measures have been proposed, each with varying degrees of predictive power. A more comprehensive understanding of these indicators would help generalise findings and offer guidance for practitioners regarding the most informative sentiment data for volatility forecasting.

Focus on investor type segmentation and regime-dependency: Another underexplored area is the differentiation of sentiment by investor type and regime. Many sentiment indices aggregate responses across groups and times, even though investors may influence volatility through different channels and time horizons. For instance, retail investors may create disruptions in the short term, whereas the sentiment from institutional investors could reflect deeper economic views and affect volatility over longer periods. To our knowledge, few studies have systematically examined whether separating sentiment by type and regime would improve volatility forecasting.

2.2.4 Contribution of This Paper

In summary, prior research has consistently shown that sentiment is linked to stock market volatility and possesses predictive power. This paper aims to address the identified research gaps by conducting a comprehensive analysis of sentiment indices that may be useful in forecasting the volatility of the S&P 500. Inspired by the framework of Paye (2012), who evaluated macroeconomic predictors and found limited improvements in forecasting accuracy, this study tests whether sentiment-based predictors offer stronger explanatory power. In doing so, it contributes to a more nuanced understanding of sentiment measures and their combined volatility forecasting performance.

The following hypotheses are, therefore, explored in this paper:

H1. Sentiment-augmented HAR models outperform traditional forecast models of realised equity market variance, both in- and out-of-sample.

H2. The predictive power of sentiment for volatility will vary by sentiment type (investor, public, institutional) and regime.

3 Data and Pre-Estimation Checks

The following section presents all data used for the analysis in this paper. It begins with exploring the dependent and independent variables included in model specifications, describing the data preparation process, and concludes with pre-estimation data checks.

3.1 Dataset Description

3.1.1 Dependent Variable

Daily realised variance, RV_t, constructed from 5-minute intraday returns, following Andersen and Bollerslev (1998), serves as the main target variable in our forecast regressions. In line with Corsi (2009), realised variance is used in contrast to volatility due to its additive and consistent nature in a linear OLS-type regression. High-frequency aggregation captures the full quadratic variation of prices and is now regarded as the benchmark measure of ex-post volatility (Andersen et al, 2003). We analyse its log-transform, $\log RV_t$, to stabilise the distribution and attenuate the right-hand tail. Figure A.1 confirms an approximately symmetric, bell-shaped histogram centred near -10, while Figure A.2 shows the well-known volatility-clustering pattern.

The realised series is taken from the Oxford-Man Institute Realised Library (Heber et al, 2009) for the S&P 500. To embed the heterogeneous-autoregressive (HAR) structure of volatility, five modifications are utilised:

$$\left\{ \log RV_{t+1},\ \log RV_t,\ \log RV_{t-1},\ \overline{\log RV}_{t:t-4},\ \overline{\log RV}_{t:t-9},\ \overline{\log RV}_{t:t-21} \right\},$$

denoted $\log RV_{\text{lead}}$, $\log RV$, $\log RV_{\text{lag}}$, $\log RV_{\text{lag\_week}}$, $\log RV_{\text{lag\_biweek}}$, and $\log RV_{\text{lag\_month}}$, respectively. This extension to the Base-HAR(1, 5, 22) is an improved fit for index-specific analysis, (Corsi, 2009), and visually matches the persistence evident in the time-series A.2: elevated volatility during the 2008–2009 global financial crisis, secondary surges in the 2011 sovereign debt crisis, the 2015–2016 Renminbi (RMB) devaluation, and the late-2018 Fed tightening episode, contrasted with "volatility-drought" intervals such as 2013–2014 and mid-2017.

3.1.2 Motivation for Predictor Selection

This thesis examines a broad set of sentiment variables to assess both their individual and collective predictive power of the S&P 500 volatility. The aim is to understand whether there are differences in how different types of sentiment predict stock market volatility. Given the behavioural foundations and the breadth of variables explored in previous literature, the selection process for this study was informed by the approach taken by Paye (2012) in his analysis of macroeconomic variables. Rather than limiting the scope to a small number of indicators, we chose to examine a wider range of variables, all supported by prior academic literature.

For this analysis, we have structured the independent variables into three dimensions:

- *Investor Sentiment:* Reflects the position of investors relative to the market, as well as their forward-looking expectations.
- *Public Sentiment:* Captures the outlook of households and businesses regarding market optimism, as well as behaviours reflecting underlying sentiment.
- *Institutional Sentiment:* Represents broader market expectations, focusing on sentiment related to macroeconomic uncertainty.

These three groups form the core of the analysis in this thesis. The groups are grounded in behavioural finance and macro-finance theory, both of which offer theoretical support for the relationship between sentiment and market volatility.

Investor sentiment - whether from retail or institutional actors - has been linked to changes in market volatility through different behavioural channels. Barberis, Shleifer and Vishny (1998), along with Barber and Odean (2008), describe how retail investors often trade excessively due to sentiment-driven biases, which can lead to mispricing and, as a result, increase volatility. On the institutional side, Shleifer and Vishny (1997) argue that capital constraints and strategic positioning can also shape behaviour. In periods of market stress, even rational investors may contribute to volatility as they adjust to shifting risk appetites and performance pressures.

Public sentiment captures the public's expectations and optimism about the future performance of the economy. Carroll et al (1994) looked into the relationship between sentiment and behaviour, finding that consumer sentiment can help predict consumption trends - suggesting that people's mood concerning the economy can affect tangible occurrences. A similar idea appears in De Bondt and Thaler (1985), who show that psychological biases among the public may result in systematic mispricings in financial markets. These insights provide a basis for considering public sentiment as an independent category in this analysis since behavioural patterns can shape how people react to economic conditions and market events.

Institutional and systemic sentiment reflects broader macroeconomic uncertainty, including asymmetries in information and the impact of exogenous shocks. A theoretical foundation for this was laid by Kyle (1985), who explained how interactions between informed traders and uninformed noise traders contribute to market volatility in the presence of uncertainty. This concept was further extended by Pástor and Veronesi (2013), who demonstrated that political and policy-related uncertainty increases risk premia and thereby volatility in the stock market. These findings provide strong theoretical support for the inclusion of systemic sentiment proxies in the analysis.

The proposed grouping is designed to capture sentiment from multiple perspectives, both to achieve a holistic understanding of its effect on volatility and to distinguish which behaviours from which groups capture and predict volatility to the largest extent. Although these three categories help structure the analysis, it is important to note that they are not mutually exclusive as all sentiment in how it affects pricing ultimately acts through investor behaviour. However, the distinction helps to derive whether signals from different sources can offer distinct predictive value.

The independent variables were selected based on the following criteria:

1. Support from prior academic literature.

2. Theoretical relevance to the relationship between sentiment and market volatility.

3. Data availability from the early 2000s to the present.

4. Compatibility with the selected modelling methodology.

This approach enables a systematic evaluation of sentiment variables and their contribution to forecasting S&P 500 volatility while ensuring consistency with established practices in prior literature. The selected predictors and their academic foundations are presented in the following section.

3.1.3 Independent Sentiment Variables

To rigorously evaluate the predictive power of sentiment on stock market volatility, several independent variables have been selected, all grounded in prior research. These variables are clustered into categories to enhance clarity and facilitate meaningful comparisons across types. The following variables serve as independent predictors in the regression analysis. Sources for the variables can be found in the Data Sources section of the References list.

Investor Sentiment Variables

Put/Call Ratio (`INV_PCSPX`): The put/call ratio (PCR) measures the daily relative demand for downside protection (put options) compared to upside speculation (call options). It reflects overall market sentiment, where a higher ratio indicates bearish expectations and a lower ratio signals bullishness. This behavioural positioning has been shown to correspond with future changes in volatility, with higher put activity coinciding with uncertainty and thereby volatility. Prior academic research, such as Gang et al (2020), demonstrates that the PCR has an asymmetric predictive relationship with stock market volatility where extreme values can signal upcoming volatility swings. In addition, Pan and Poteshman (2006) find that the PCR has predictive power over stock market prices, which are closely linked to volatility. These findings provide a strong basis for including the PCR as an independent variable in the analysis.

CBOE SKEW Index (`INV_SKEW`): The CBOE SKEW index captures how investors perceive the chance of a significant downside event - commonly referred to as a "black swan" - in the S&P 500 over the next 30 days. When the SKEW rises, it is often viewed as a sign that concerns about a market crash are increasing, pointing to growing uncertainty among investors. Past research links said uncertainty to increased volatility. Mora-Valencia et al (2021) found that the SKEW, both by itself and in combination with the VIX, could help predict stock market volatility. While the VIX reflects expected variance, the SKEW more closely reflects the possibility of extreme shifts in the tail of the distribution. While SKEW is often included as a sentiment variable it may also reflect more fundamental pricing of rare-event probabilities. We acknowledge this dual interpretation but chooses to follow sentiment literature like Mora-Valencia et al (2021) and include it as a sentiment variable reflecting fear-based option demand.

AAII Investor Sentiment Survey (`INV_BULL; INV_BEAR`): The weekly survey conducted by the American Association of Individual Investors captures whether retail investors are bullish, bearish, or neutral regarding the stock market over the coming six months. This variable, therefore, reflects the sentiment of individual, noninstitutional investors regarding their expectations about future market performance. Kresta et al (2024), for instance, analyse the predictive power of the AAII Sentiment Survey for stock market volatility and find a significant negative relationship: optimism among retail investors tends to correspond with calmer markets, while elevated fear is associated with future turbulence. This variable captures sentiment from the perspective of mainstream investors and complements other measures by focusing specifically on individual investor outlooks. Its inclusion allows us to test the predictive power of individual retail investors on stock market volatility.

NAAIM Exposure Index and its Standard Deviation (`INV_EXPOSURE; INV_EXPOSURE_STD`): The NAAIM Exposure Index, published by the National Association of Active Investment Managers, measures the weekly average equity exposure of active money managers. Higher values reflect optimism and lower values suggest more defensive positioning. The standard deviation of this index captures the level of disagreement among these money managers regarding their market outlook. A high standard deviation implies divergent views and a lack of consensus, whereas a low value suggests alignment in positioning. Academic literature supports both measuring the index level and its dispersion as informative sentiment measures. The fundamental theory suggests that widespread optimism may correspond with lower volatility. However, Ben-David et al (2018) show that ETF ownership, used heavily by institutional investors, can increase volatility. This suggests that excessive market exposure can amplify volatility, especially when many actors pursue similar strategies. Additionally, Hong and Stein (2007) argue that a lack of consensus among investors can lead to price inefficiencies, which may also trigger spikes in volatility. These findings justify the inclusion of both the index and its standard deviation in the model.

Bullish Percentage Index (`INV_BPI`): The Bullish Percentage Index (BPI) is a technical sentiment indicator that reflects how broadly bullish signals appear across the stocks within a given index. Rather than focusing on individual price movements, the BPI captures the extent of upward momentum within the market as a whole. A high reading tends to coincide with overall optimism, while a low BPI points toward widespread pessimism - both of which may be associated with changes in volatility. Some behavioural theories argue that sentiment at extreme levels - whether overly positive or negative - can lead to exaggerated market reactions and greater price swings. Lee et al (2002) examined sentiment indicators similar to the BPI and observed that volatility tends to rise in both highly optimistic and highly pessimistic environments. This suggests a nonlinear effect, where extremes in sentiment may create greater uncertainty. These findings provide support for using the BPI as a relevant sentiment proxy in the model.

Margin Debt (`INV_MARGINDEBT`): Margin debt refers to the amount of money investors have borrowed to purchase stocks on margin. This measure thus reflects leveraged investor positioning and can serve as a proxy for the risk appetite of investors as well as financial leverage. The academic literature is somewhat mixed in the connection between margin debt and volatility, as summarised by Fortune (2001). However, he concludes that margin debt does influence stock market volatility, although the effect is relatively limited. Despite inconclusive empirical evidence, the variable is included in the analysis due to its theoretical relevance as an indicator of speculative market behaviour.

Public Sentiment Variables

University of Michigan Consumer Sentiment Index (`CON_SENT`): The University of Michigan Consumer Sentiment Index is used to track how the public views the broader state of the economy. While it is not a direct measure of stock market expectations, the index reflects general confidence in future conditions, which may correlate with broader market sentiment. Garcia and Carvalho (2025) show that stronger consumer sentiment is typically linked to lower stock market volatility, indicating a possible stabilising effect. Similarly, Sum and Chorlian (2013) find that consumer confidence helps explain variation in stock market risk premiums, supporting the idea that household sentiment has an influence on financial market dynamics, including volatility. Based on this, the index can be seen as a relevant proxy for public mood and its potential link to volatility.

New Business Applications (`CON_NEW_BUSINESS`): The weekly rate of new business applications in the United States, reported by the Census Bureau, serves as a proxy for entrepreneurial and business confidence. The rationale for including this variable lies in the idea that the number of new businesses reflects optimism about the future state of the economy and a willingness to take on risk. Bonato et al (2024) support this view by analysing the predictive value of business applications to stock market volatility in the U.S., finding a significant result over intermediate and long horizons, even after controlling for various realised moments in the HAR model. Including this variable allows the model to capture a broader definition of investor sentiment, relevant to the implications of the thesis.

Fannie Mae Refinance Application-Level Index (`CON_REMORTGAGE`): The Fannie Mae Refinance Application-Level Index reflects the weekly volume of mortgage refinance applications in the U.S. The intuition behind this variable is that when households are confident about the economic outlook, or when interest rates are low, people refinance their mortgages to capture superior terms. As such, the index represents household optimism about personal finances and broader financial conditions. Although limited research has directly examined the relationship between mortgage financing and stock market volatility, Mian and Sufi (2011) link refinancing not only to economic fundamentals but also to sentiment-driven behaviours such as optimism and confidence. Additionally, Keys et al (2014) demonstrate that refinancing predicts shifts in consumption and investment behaviour, suggesting that it can amplify financial fluctuations. These insights support the inclusion of the refinancing index as a sentiment-related predictor in the model.

Institutional Sentiment Variables

Economic Policy Uncertainty Index (`INST_POLICY_UNCERTAINTY`): The Economic Policy Uncertainty (EPU) Index tracks how often the term "economic policy uncertainty" appears in newspaper articles. Baker, Bloom, and Davis (2016) introduced it as a tool for gauging the level of uncertainty tied to government policy and its potential influence on investor behaviour. When uncertainty increases, it can raise concerns among investors, leading to higher risk premiums, changes in option pricing, and potentially more volatile markets. Liu and Zhang (2015) find support for this idea in their work, showing that greater policy-related uncertainty tends to coincide with increases in volatility. Given this relationship, the EPU Index is treated as a relevant input in this analysis.

Monetary Policy Uncertainty Index (`INST_MONETARY_UNCERTAINTY`): The Monetary Policy Uncertainty (MPU) Index, developed as an extension of the EPU index, captures uncertainty related to U.S. monetary policy. It is constructed by tracking the frequency of monetary policy terms appearing in newspaper articles. High values indicate greater uncertainty regarding actions by the Federal Reserve, which may influence stock market volatility through changes in interest rate risk and a general reduction of clarity in the market. Academic literature gives a nuanced picture of the index, where Hsiao et al (2022) find that a high MPU, contrary to theoretical expectations, negatively affects stock market volatility, even when controlling for other economic variables. This counterintuitive result underscores the complexity of the index's effects, yet its demonstrated predictive power justifies its inclusion as an independent variable in the analysis.

Presidential Approval Rating (`INST_PRES_APP`): The approval rating of the sitting U.S. president reflects general public sentiment toward institutional trust. Declining approval may indicate rising uncertainty regarding economic stability as well as regulations and general policies, contributing to increased volatility. Gupta et al (2023) demonstrate that presidential approval ratings have a statistically significant out-of-sample predictive power for stock market volatility in the U.S. This relationship can be attributed to the fact that higher approval ratings often signal political stability, fostering calmer markets. As such, the rating captures volatility through a summary of the socio-political climate and is included to account for public trust.

Geopolitical Risk Index (`INST_GEOPOL_RISK`): The Geopolitical Risk (GPR) Index is constructed by quantifying the frequency of geopolitical tensions being discussed in newspaper articles. This measure captures global uncertainty, relating to fear within investor networks, triggering market reactions. Caldara and Iacoviello (2018), who developed this index, find that higher GPR correlates with an increased downside risk and, consequently, increased volatility. Including this sentiment measure helps ensure that more global fear factors are included that may be overlooked by domestic news channels.

FRBSF Daily News Sentiment (`INST_NEWS_SENT`): The FRBSF Daily News Sentiment, developed by the Federal Reserve Bank of San Francisco, quantifies sentiment in economics-related news articles using a text analysis to indicate whether the discussions are positive or negative. In theory, this reflects an overall rising or falling concern about the economic outlook, which is expected to correlate with changes in market volatility. Macro news sentiment has previously been linked to predicting volatility, for instance, by Bodilsen and Lunde (2024), who found substantial improvement in accuracy in their predictive models for volatility by incorporating macro news sentiments. Although the FRBSF is a more novel approach, it is empirically supported as a relevant predictor for volatility.

TED Spread (`INST_TED`): The TED Spread represents the gap between the 3-month LIBOR rate and the 3-month U.S. Treasury bill yield. This difference tends to widen when banks become more reluctant to lend, and it has been linked to broader concerns in the financial system. Though originally used to monitor banking sector stress, it is also considered relevant for assessing general market conditions. A rising TED Spread can indicate growing risk aversion, and has, in some cases, been followed by increased volatility. Mittnik et al (2015) discuss how sharp movements in the spread have coincided with turbulent market periods, particularly during liquidity shortages. For that reason, the TED Spread is included as one of the explanatory variables in this model.

High-Yield Bond Spread (`INST_HYBS`): The High-Yield Bond Spread compares the yield on lower-rated corporate bonds to that of equivalent-maturity U.S. Treasuries. It tends to widen during periods when investors grow more cautious and demand greater compensation for taking on credit risk. This shift is often taken as a sign of market stress and has been linked to changes in expected volatility. In their study, Chun et al (2023) identified the spread as one of the few variables with a significant predictive value under the HAR framework. For this reason, the High-Yield Bond Spread is used in the model as a sentiment-related indicator grounded in credit market dynamics.

Control Variables

Volatility Index ($logVIX$): The VIX is one of the most typical sentiment-based predictors of stock market volatility, derived from the S&P 500 option prices. It is therefore included as a control variable, as it directly reflects how investors predict future volatility and provides a forward-looking benchmark. This allows us to assess the incremental predictability of the other independent variables, beyond what is captured by the VIX. Empirical studies, such as Blair, Poon and Taylor (2000) and Chen and Li (2021), have shown that the VIX statistically predicts realised variance. Its inclusion ensures that the forecasting performance of additional variables is evaluated relative to this well-established benchmark. The natural logarithm is applied to the variable to ensure consistency with the dependent variable.

Daily Return ($daily\_return$): The daily return, calculated by the percentage change in the S&P 500 index from one day to the next, is included as a control variable to account for the leverage effect. This phenomenon, first explained by Black (1976), suggests that when stock prices fall, firms tend to become more leveraged, increasing their risk and, in turn, volatility. Including the daily return helps improve model accuracy by capturing the volatility clustering behaviour observed in financial markets.

Jump Component ($jump$): The jump component is made as a binary variable and captures large, discrete price movements that deviate from the usual return distribution. These events, while rare, can have a significant impact on realised variance and are often associated with unexpected news or macroeconomic shocks. In line with existing literature on the importance of accounting for jumps (Barndorff-Nielsen and Shephard, 2004; Bollerslev and Todorov, 2011), we include a simplified approximation of jump risk. Specifically, we define a day as a "jump day" if the realised variance falls within the top 5% of all previous daily realised variances in the sample and is calculated as $I(logRV > 95thPercentile)$. This percentile-based binary method allows us to flag extreme movements in a computationally efficient way and capture some of the tail risks that traditional continuous-volatility measures might miss.

3.2 Data Preparation

Data preparation procedures employed before model estimation and forecasting analyses are as follows. The initial dataset comprises variables sampled at varying frequencies - daily, weekly, and monthly. Weekly and monthly variables are forward-filled to align these frequencies and maintain the temporal integrity required for daily volatility prediction within the HAR framework. Specifically, recent observations are propagated forward to fill subsequent daily observations until the following observation. Such forward-filling allows for analysis at the daily level while mimicking investors' information sets and ensuring no look-ahead bias contaminates the forecasts (Paye, 2012). This study uses contemporaneously measured sentiment indicators to examine their predictive power for volatility within the same month. In practice, however, many of these indicators - particularly survey-based or media-derived measures (see e.g. Baker et al, 2016) - are published with a delay. As such, the resulting forecasts can be interpreted as an upper bound on predictive performance, reflecting a hypothetical scenario in which sentiment data are available in real time. While this limits the immediate practical applicability of the models, the approach remains empirically meaningful for evaluating the role of sentiment in volatility prediction.

Furthermore, the final dataset is carefully aligned by restricting the analysis to common data availability across all included variables. Consequently, the effective time frame for analysis is constrained by the variables with the shortest available time series. `INV_PCSPX` begin on 18/10/2006, which, hence, becomes our starting date. $logRV$ is obtained through the Oxford-Man Realised Variance dataset, which terminated in 2020, and hence, data is not available after 31/03/2020. We note that this coincides with the start of COVID-19 which could affect the results, but as our sample covers a long time period we consider it acceptable. Furthermore, weekends and national holidays are removed as daily data in both dependent and independent variables are not available. This approach avoids bias from missing or misaligned observations and is consistent with standard empirical methods (Andersen et al, 2003; Corsi, 2009).

Additionally, all sentiment and uncertainty variables included in the analysis undergo standardisation, which serves several critical purposes within this research context and follows common practice in volatility forecasting (e.g. Paye, 2012). Firstly, it ensures that all predictors enter the models on comparable scales, eliminating potential biases arising from scale differences such as multicollinearity and parameter instability while estimating the predictors. Comparability is essential when implementing regularisation techniques like LASSO or dimension-reduction methods such as PCA, where variables of larger scale might otherwise disproportionately influence the penalisation or factor extraction processes (Hastie et al, 2009). Secondly, standardisation facilitates an intuitive comparison between estimated coefficients. Their respective importance in predicting volatility can then be directly compared. To further ensure comparability, the VIX is log-transformed prior to standardisation to align with the scale of the dependent variable.

Together, these preparation steps lay the foundation for this work's analysis, providing the following dataset.

3.3 Summary Statistics and Data Visualisation

Table 3.1: Descriptive Statistics

	Mean	SD	Q1	Median	Q3	Min	Max	Skewness	Kurtosis
logRV	-10.02	1.25	-10.90	-10.16	-9.25	-13.62	-4.86	0.52	0.37
logRV_lag	-10.02	1.25	-10.90	-10.16	-9.25	-13.62	-4.86	0.52	0.37
logRV_lag_week	-10.03	1.15	-10.83	-10.14	-9.35	-12.82	-5.80	0.63	0.51
logRV_lag_biweek	-10.03	1.10	-10.80	-10.17	-9.39	-12.47	-6.11	0.68	0.55
logRV_lag_month	-10.04	1.04	-10.74	-10.23	-9.44	-12.25	-6.35	0.71	0.59
jump	0.05	0.23	0.00	0.00	0.00	0.00	1.00	3.93	13.46
daily_return	0.00	0.01	0.00	0.00	0.01	-0.12	0.12	-0.28	14.21
logVIX	2.87	0.39	2.59	2.80	3.09	2.21	4.42	1.03	1.15
INV_SKEW	0.32	1.00	-0.41	0.16	0.89	-1.84	4.58	0.74	0.33
INV_PCSPX	0.00	1.00	-0.58	-0.14	0.39	-2.25	18.59	4.68	55.32
INV_BPI	0.12	1.01	-0.48	0.45	0.86	-3.08	1.64	-0.99	0.43
INV_BULL	-0.26	0.80	-0.85	-0.29	0.30	-2.14	2.42	0.17	-0.30
INV_BEAR	0.14	1.00	-0.63	0.16	0.84	-2.70	3.03	0.03	-0.43
INV_EXPOSURE	0.02	1.00	-0.59	0.21	0.79	-2.83	2.35	-0.65	-0.29
INV_EXPOSURE_STD	0.00	1.00	-0.75	0.02	0.68	-2.63	2.70	0.04	-0.47
INV_MARGINDEBT	0.03	1.15	-0.54	0.22	0.77	-5.38	4.11	-1.14	2.55
CON_SENT	-0.27	1.03	-1.07	-0.26	0.69	-2.36	1.17	-0.33	-1.13
CON_NEW_BUSINESS	0.01	1.02	-0.60	-0.05	0.60	-4.39	3.27	0.04	1.20
CON_REMORTGAGE	0.01	1.07	-0.40	-0.04	0.33	-4.80	8.32	1.65	11.28
INST_HYBS	0.01	1.09	-0.48	-0.01	0.37	-9.52	13.64	1.41	21.87
INST_TED	0.07	1.16	-0.54	-0.30	0.05	-0.84	10.28	3.51	16.61
INST_POLICY_UNCERTAINTY	0.28	1.02	-0.43	0.02	1.00	-1.79	3.73	0.64	0.19
INST_GEOPOL_RISK	-0.23	0.32	-0.45	-0.28	-0.08	-0.77	0.88	1.16	1.38
INST_MONETARY_UNCERTAINTY	-0.15	0.85	-0.76	-0.35	0.26	-1.17	3.74	1.67	3.74
INST_NEWS_SENT	0.00	1.02	-0.60	-0.04	0.53	-6.27	5.41	0.26	2.98
INST_PRES_APP	-0.37	0.65	-0.77	-0.34	-0.08	-1.98	1.57	0.10	0.42

Descriptive Statistics

Table 3.1 reports summary statistics on the complete dataset, with 3,511 observations dated 18/10/2006 to 31/03/2020. The summary includes the dependent variable - daily log realised variance of the S&P 500 - and the observed HAR, control, and sentiment regressors. By construction, all explanatory variables have means close to zero and their standard deviations close to one.

The reported mean, skewness, and kurtosis of realised variance show the expected benefit of logging the dependent variable. Namely, the variable has become more linear-model friendly with its increased symmetry and reduced tails. Nonetheless, some extreme observations remain: the minimum of –13.62 coincides with the unusually tranquil summer of 2017, whereas the maximum of –4.86 highlights the height of the global financial crisis in late 2008.

The controls mostly match the distribution expected from such a variable where the jump and daily return show high levels of kurtosis due to extreme observations in the S&P 500. This is consistent with prior evidence showing fat tails and excess kurtosis in return-based variables (Andersen et al, 2003), echoing foundational insights from Mandelbrot (1963) and Fama (1965).

The sentiment covariates all live on a comparable, standardised scale, with some showing high skew and kurtosis due to their intrinsic asymmetric shape. For example, the put/call ratio of the S&P 500 exhibits sharp, narrow spikes of extreme deviation during financial crises, whilst heavily clustered around the mean during other periods. Moments of shock-induced financial volatility in the dataset help produce such kurtosis in the variables of interest (Kelly and Jiang, 2014; Jena et al, 2019).

Figure A.3 summarises the cross-sectional distribution of the 18 standardised sentiment and uncertainty series. Because each variable was demeaned and scaled to unit variance before transformation, the central medians cluster tightly around zero. Inter-quartile ranges vary, however, indicating heterogeneous volatility of the underlying signals where consumer sentiment and policy uncertainty display the widest boxes, whereas the TED rate and geopolitical risk are markedly more compact. Several indices exhibit material skew and tail risk even after the signed $\log(1 + |x|)$ re-scaling - most visibly the positive tails of the put/call options ratio, consumer remortgages, and high yield bond savings - suggesting episodic spikes that could carry predictive information for volatility jumps.

Stationarity Tests

In time series econometrics, testing for and achieving stationarity is a critical step before model estimation, especially when working with predictive frameworks like the Heterogeneous Autoregressive (HAR) model. Nonstationary data - characterised by a time-varying mean, variance, or autocovariance - can lead to spurious regression results, thereby invalidating statistical inference (Granger and Newbold, 1974; Dickey and Fuller, 1979; Hamilton, 1994). This is particularly relevant in financial and economic applications, where many macroeconomic and market variables inherently exhibit trending behaviour.

In line with this, several variables in our dataset were found to be nonstationary - specifically, Investor Margin Debt, High Yield Bond Savings, Consumer Sentiment Index and Presidential Approval Ratings. To transform these into stationary series, we applied seasonal differencing tailored to the natural frequency of each variable. High Yield Bond Savings (HYBS), being sensitive to high-frequency institutional activity, was differenced at a daily lag (1-day). Investor Margin Debt, Consumer Sentiment Index, and Presidential Approval Ratings - typically updated or evolving more gradually - were differenced using a monthly lag (22 trading days) to remove persistent seasonality and trend components without over-differencing. These differencing choices are consistent with standard transformations used in financial time series, where one month is approximated as 22 trading days (Tsay, 2005; Zivot and Wang, 2006), and seasonal differencing is commonly employed to remove periodic trends and achieve stationarity (Enders, 2014).

Table 3.2: ADF Stationarity Test Results

Variable	ADF Statistic	p-value	Lag Used	Stationary (p < 0.05)
logRV	-5.801	< 0.010	15	Yes
logRV_lead	-5.829	< 0.010	15	Yes
logRV_lag	-5.720	< 0.010	15	Yes
logRV_lag_week	-4.310	< 0.010	15	Yes
logRV_lag_biweek	-5.800	< 0.010	15	Yes
logRV_lag_month	-5.246	< 0.010	15	Yes
daily_return	-15.601	< 0.010	15	Yes
jump	-6.909	< 0.010	15	Yes
logVIX	-3.834	0.0172	15	Yes
INV_SKEW	-5.786	< 0.010	15	Yes
INV_PCSPX	-9.819	< 0.010	15	Yes
INV_BPI	-6.454	< 0.010	15	Yes
INV_BULL	-7.163	< 0.010	15	Yes
INV_BEAR	-4.729	< 0.010	15	Yes
INV_EXPOSURE	-5.787	< 0.010	15	Yes
INV_EXPOSURE_STD	-8.616	< 0.010	15	Yes
INV_MARGINDEBT	-9.198	< 0.010	15	Yes
CON_SENT	-6.545	< 0.010	15	Yes
CON_NEW_BUSINESS	-7.019	< 0.010	15	Yes
CON_REMORTGAGE	-10.345	< 0.010	15	Yes
INST_HYBS	-11.826	< 0.010	15	Yes
INST_TED	-4.761	< 0.010	15	Yes
INST_POLICY_UNCERTAINTY	-3.825	0.0177	15	Yes
INST_GEOPOL_RISK	-8.097	< 0.010	15	Yes
INST_MONETARY_UNCERTAINTY	-5.716	< 0.010	15	Yes
INST_NEWS_SENT	-13.316	< 0.010	15	Yes
INST_PRES_APP	-4.750	< 0.010	15	Yes

Inter Correlations and Multicollinearity Considerations

High pairwise correlations among some sentiment proxies raise potential multicollinearity concerns (Table A.1). For instance, `INST_MONETARY_UNCERTAINTY` correlates 0.59 with `INST_POLICY_UNCERTAINTY`. Similarly, investor positioning metrics - `INV_EXPOSURE` and `INV_BPI`- exhibit correlations above 0.52, and `INV_MARGINDEBT` and `INV_BPI` exhibit correlations above 0.53. (See Figure A.4 for the correlation matrix of predictors).

To assess the extent to which multicollinearity may compromise the precision of coefficient estimates, Variance Inflation Factors (VIFs) are computed for all explanatory variables. Following established guidelines, VIF values exceeding 5 are typically considered indicative of problematic multicollinearity, whereas values near 1 suggest negligible correlation among predictors (James et al, 2021). As reported in Table A.2, the VIFs associated with the sentiment-based regressors range from 1.06 to 4.15, suggesting a moderate correlation that does not preclude reliable statistical inference. Nonetheless, to further address potential redundancy among covariates, Least Absolute Shrinkage and Selection Operator (LASSO) regularisation is explored, as it systematically penalises overfitting and attenuates the influence of collinear predictors (Tibshirani, 1996).

4 Methodology

4.1 Model Framework

4.1.1 Overview

To assess whether sentiment-based indicators improve the predictive accuracy of daily realised variance, this thesis employs a structured sequence of econometric models, each incrementally expanding the information set. The incremental approach facilitates isolating specific contributions to forecasting performance, with four primary specifications forming the core analysis, and two additional models offering complementary insights.

The analysis begins with the Heterogeneous Autoregressive (HAR) model, which serves as the basis for forecasting next-day volatility solely based on its historical components (Base-HAR). We employ a HAR framework over GARCH-family models due to its straightforward integration of mixed-frequency sentiment data, that does not impose restrictive functional forms or distributional assumptions characteristic of traditional GARCH approaches. An extended version of the HAR model establishes the baseline level of predictive accuracy achievable using conventional data by incorporating control variables and a widely used benchmark for volatility forecasting (HAR). Following this, HAR-X augments this baseline with a comprehensive set of sentiment indicators to evaluate whether forward-looking sentiment contains incremental predictive power beyond that embedded in past volatility. A theoretical refinement of this framework is the HAR-X-LASSO model, which retains the complete set of sentiment covariates but introduces a regularisation penalty, enhancing model parsimony and mitigating multicollinearity. Together, these four specifications - Base-HAR, HAR with controls, HAR-X, and HAR-X-LASSO - form the core of the empirical comparison, enabling a structured examination of the relative contributions of historical volatility, sentiment content, and shrinkage-based regularisation to forecast performance.

Two additional models offer complementary insights. The HAR-X-PCA compresses the sentiment block into a few principal components, allowing for dimensionality reduction and interpretation of sentiment-driven forecasting through the lens of latent factors. Lastly, the Jump-HAR-X model augments the leading specification with a jump component to capture discontinuous shocks in volatility, thereby testing the capacity of sentiment variables to anticipate sudden, nonlinear shifts in market conditions.

Together, this ordered sequence provides a comprehensive assessment of the channels through which sentiment may contribute to volatility forecasting.

4.1.2 Heterogeneous Autoregressive (HAR) Model of Realised Variance

A model that has gained substantial prominence in literature for its simplicity and accuracy in capturing the long-memory characteristics of volatility is the Heterogeneous Autoregressive (HAR) model, initially introduced by Corsi (2009). Unlike classical GARCH-based frameworks, the HAR model leverages realised variance RV measures derived from high-frequency intraday returns, providing a more precise and direct assessment of the latent volatility process. In line with Corsi (2009), this work extends the baseline HAR(1,5,22) model to HAR(1,2,5,10,22), a simple yet highly effective autoregressive structure to model index volatility. Expressed formally:

$$\begin{aligned}\log(RV_{t+1}) = \beta_0 + \beta_d \log(RV_t^{(d)}) + \beta_l \log(RV_{t-1}^{(l)}) + \beta_w \log(RV_t^{(w)}) \\ + \beta_b \log(RV_t^{(b)}) + \beta_m \log(RV_t^{(m)}) + \gamma r_t + \mu j_t + \delta \log(VIX_t) + \varepsilon_{t+1}\end{aligned} \tag{4.1}$$

where $\log(RV_{t+1})$ is the logarithm of realised variance on day $t+1$, and the explanatory variables include lagged realised variance terms defined over different horizons: the daily realised variance today, $\log(RV_t^{(d)})$, and yesterday, $\log(RV_{t-1}^{(l)})$, weekly realised variance $\log(RV_t^{(w)}) = \frac{1}{5}\sum_{i=0}^{4}\log(RV_{t-i})$, bi-weekly realised variance $\log(RV_t^{(b)}) = \frac{1}{10}\sum_{i=0}^{9}\log(RV_{t-i})$, and monthly realised variance $\log(RV_t^{(m)}) = \frac{1}{22}\sum_{i=0}^{21}\log(RV_{t-i})$.

This model specification, and with the use of ordinary-least-squares estimation, allows all coefficients to be interpreted in terms of percentage change in realised variance, providing clear financial intuition. In the case of the baseline HAR structure, this extends to elasticities. In particular, β_d and β_l are expected to capture the short-run dynamics in variance, while β_w, β_b, and β_m reflect intermediate and long-run persistence respectively. Each coefficient represents the percentage change in realised variance, given a one percentage change in the variable.

Additionally, daily return is included to account for the well-documented leverage effect, wherein negative returns are associated with disproportionately higher future volatility compared to positive returns of equivalent magnitude (see Black, 1976; Christie, 1982). Formally, for each one-unit increase in daily return (1% increase in returns), a γ% change is expected in realised variance. A jump component is also incorporated as a control to take into account discontinuous price movements, capturing the impact of abrupt financial shocks that cannot be explained by the continuous volatility process alone (Barndorff-Nielsen and Shephard, 2004). Given its binary nature, we can expect periods of extreme volatility to be associated with a μ % change in realised variance.

Finally, the CBOE Volatility Index ($log(VIX)$) - a widely recognised forward-looking measure derived from option-implied volatility - is included as a contemporaneous control. Given its market-based nature, the VIX serves as a natural benchmark for expected future volatility and provides a point of comparison for the incremental predictive value of sentiment-based indicators.

4.1.3 HAR-X Model with Sentiment Variables

While the baseline HAR model (Corsi, 2009) offers a robust approach for forecasting realised variance, recent literature has increasingly emphasised incorporating exogenous information into volatility modelling (e.g., Verma and Verma, 2007; Kelly and Jiang, 2014; Patton and Sheppard, 2015; Mittnik et al, 2015; Audrino et al, 2020; Ding et al, 2021; Bai et al, 2024; Li et al, 2024). As such, this thesis extends the baseline HAR framework into a HAR-X model, a specification that integrates various exogenous regressors. Formally, the HAR-X model adopted can be expressed as follows:

$$\log(RV_{t+1}) = \beta_0 + \beta_d \log(RV_t^{(d)}) + \cdots + \sum_{j=1}^{p} \theta_j S_{j,t} + \varepsilon_{t+1} \tag{4.2}$$

where the sentiment regressors $S_{j,t}$ are included contemporaneously (day t) to ensure no look-ahead bias occurs in the estimation procedure. Specifically, the sentiment variables incorporated in the analysis include a comprehensive set of investor, public and institutional sentiment proxies. We anticipate specific sign conventions for the associated coefficients (θ_j): negative shocks to sentiment (i.e., increased pessimism or fear captured by higher economic policy uncertainty or geopolitical risk) are generally expected to predict higher subsequent volatility, resulting in positive coefficients. Conversely, increased optimism - such as improved consumer sentiment or bullish investor sentiment - is expected to lead to lower future volatility, thus associated with negative coefficients.

To investigate whether the influence of sentiment on future volatility is state-dependent, we further develop a regime-based extension of the HAR-X framework. Specifically, we introduce a dynamic threshold HAR-X model, where the impact of each sentiment variable is permitted to vary nonlinearly across high and low volatility environments.

To determine an optimal volatility threshold, denoted by c_t^*, we employ a dynamic rolling-window approach. For each day t, the threshold c_t^* is selected exclusively based on all available historical data up to day $t-1$. Specifically, for each day t, we conduct a separate grid-search procedure over a range of candidate volatility thresholds:

$$c_t^* = \arg\min_{c \in C_t} \sum_{\tau=1}^{t-1} \left(\hat{\varepsilon}_\tau(c)\right)^2,$$

where $\hat{\varepsilon}_\tau(c)$ are regression residuals from estimating the model using historical data up to day $t-1$ and candidate threshold c, and C_t is the grid of candidate threshold values derived from quantiles of historical volatility data. This sequential, past-data-only approach eliminates any forward-looking bias, ensuring realistic and unbiased regime classification.

Once c_t^* is dynamically determined for each day, it partitions the sample at time t into two regimes: periods of relatively low market volatility ($\log RV_t < c_t^*$) and periods of heightened market stress ($\log RV_t \geq c_t^*$). Each sentiment regressor is then interacted with this day-specific regime indicator, yielding a dynamic piecewise-linear formulation:

$$\log(RV_{t+1}) = \beta_0 + \beta_d \log(RV_t^{(d)}) + \cdots + \sum_{j=1}^{p} \left(\theta_j^{(L)} S_{j,t}^{(L)} + \theta_j^{(H)} S_{j,t}^{(H)} \right) + \varepsilon_{t+1}, \tag{4.3}$$

where:

$$S_{j,t}^{(L)} = S_{j,t} \cdot \mathbb{1}(V_t < c_t^*), \quad S_{j,t}^{(H)} = S_{j,t} \cdot \mathbb{1}(V_t \geq c_t^*),$$

and $\theta_j^{(L)}$ and $\theta_j^{(H)}$ represent the marginal effects of sentiment variable j in low- and high-volatility regimes, respectively.

This dynamic specification allows for robust statistical testing of coefficient heterogeneity across volatility regimes via standard Wald tests on the null hypothesis $H_0 : \theta_j^{(L)} = \theta_j^{(H)}$. Intuitively, this flexible and realistic setup permits market sentiment signals to have regime-dependent impacts, reflecting that markets may "listen" differently to sentiment under calm versus turbulent conditions, without introducing look-ahead bias.

In supplementary analysis, a probit-based model of volatility jumps is included, wherein the dependent variable is a binary indicator capturing 95th percentile volatility episodes. This provides an alternative lens on whether sentiment measures help anticipate abrupt market dislocations, conditional on recent extreme variance dynamics.

$$\begin{aligned} Jump_{t+1} = \beta_0 &+ \beta_d \log(Jump_t^{(d)}) + \beta_l \log(Jump_{t-1}^{(l)}) + \beta_w \log(Jump_t^{(w)}) \\ &+ \beta_b \log(Jump_t^{(b)}) + \beta_m \log(Jump_t^{(m)}) + \gamma r_t + \mu j_t + \delta \log(VIX_t) \\ &+ \sum_{j=1}^{p} \theta_j S_{j,t} + \varepsilon_{t+1} \end{aligned} \tag{4.4}$$

4.1.4 LASSO Regularisation for Variable Selection

To address the challenges posed by incorporating a large number of sentiment and uncertainty variables - namely multicollinearity, overfitting, and reduced out-of-sample performance - this thesis employs the Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive LASSO) within the HAR-X framework. Adaptive LASSO improves upon standard LASSO by applying variable-specific penalties, enabling consistent variable selection and reducing bias in coefficient estimation (Zou, 2006).

The procedure, which follows methods described by Zou (2006) and Ding et al (2021) unfolds in two stages. First, ridge regression is applied to the sentiment predictors to obtain stable, nonzero initial estimates $\hat{\theta}_j$, which are used to construct adaptive weights:

$$w_j = \frac{1}{|\hat{\theta}_j|^{\gamma}} \quad \text{with} \quad \gamma > 0 \tag{4.5}$$

Second, these weights are applied in a weighted LASSO regression estimated via the following objective function:

$$\min_{\theta} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left(\log(RV_{t+1}) - \sum_{j=1}^{p} \theta_j \, S_{j,t} \right)^2 \; + \; \lambda \sum_{j=1}^{p} w_j \, |\theta_j| \right\} \tag{4.6}$$

where $S_{j,t}$ denotes the sentiment predictors, and λ controls the level of penalisation. The optimal λ is selected using the criterion such that the cross-validated error is within one standard error of the minimum (James et al, 2021).

The selected sentiment variables are then incorporated into a linear HAR-X model alongside unpenalised core volatility components. This two-step HAR-X-Adaptive-LASSO approach retains key volatility dynamics while allowing only the most predictive sentiment signals to enter the final specification. Compared to standard LASSO, Adaptive LASSO offers improved robustness in the presence of correlated predictors and satisfies the oracle property under certain conditions, allowing it to identify the true set of nonzero coefficients. This makes it especially well-suited for high-dimensional applications involving noisy and potentially collinear sentiment inputs.

4.1.5 Principal Component Analysis (PCA) for Dimension Reduction

To complement the primary variable selection strategy, this thesis incorporates Principal Component Analysis (PCA) into the HAR-X framework. PCA is an unsupervised dimensionality reduction technique that transforms a potentially large and correlated set of sentiment and uncertainty variables into a smaller set of orthogonal factors known as principal components.

By construction, PCA linearly re-expresses the original standardised variables $S_{j,t}$, as uncorrelated linear combinations that sequentially capture the greatest possible variance in the data. Formally, the standardised data matrix X is decomposed as:

$$X = P \Delta Q^T \tag{4.7}$$

where Q contains the component loadings, Δ is a diagonal matrix of singular values, and the transformed matrix XQ yields the principal component scores. These scores, denoted PCk, t, are then included as predictors in the following PCA-augmented HAR-X model:

$$
\begin{aligned}
\log(RV_{t+1}) = \beta_0 &+ \beta_d \log(RV_t^{(d)}) + \beta_l \log(RV_{t-1}^{(l)}) + \beta_w \log(RV_t^{(w)}) \\
&+ \beta_b \log(RV_t^{(b)}) + \beta_m \log(RV_t^{(m)}) + \gamma r_t + \mu j_t + \delta \log(VIX_t) \\
&+ \sum_{k=1}^{K} \phi_k PC_{k,t} + \varepsilon_{t+1}
\end{aligned} \tag{4.8}
$$

The number of components K is selected based on the cumulative explained variance being greater than 50%, which ensures that they include the minimal set of components that together explain a substantial proportion of total variance.

The inclusion of HAR-X-PCA allows for a compact and interpretable representation of sentiment dynamics, enabling the assessment of how composite sentiment trends relate to future volatility. PCA also provides a theoretically grounded and computationally efficient approach to incorporating high-dimensional sentiment data. It enhances the robustness and interpretability of the HAR-X framework and serves as a valuable benchmark alongside unrestricted and regularised model specifications.

4.2 Forecast Evaluation Framework

Estimation of the HAR model is straightforward using Ordinary Least Squares (OLS), supported by robust inference procedures that account for potential heteroskedasticity and autocorrelation in residuals. Specifically, following recent best practices, Newey-West heteroskedasticity and autocorrelation consistent (HAC) standard errors with appropriately large bandwidth choices are employed to ensure valid inference (Newey and West, 1987). Due to its simplicity, interpretability, and strong empirical forecasting performance, the baseline HAR model described here forms the primary benchmark against which more complex and theoretically enriched models - such as those incorporating sentiment variables and advanced regularisation methods - will be systematically evaluated in subsequent analysis.

To robustly evaluate and compare the forecasting performance of the HAR-type models described earlier, we adopt a pseudo-out-of-sample forecasting methodology, widely employed in volatility prediction literature (Corsi, 2009; Paye, 2012). In generating out-of-sample forecasts, we employ an expanding window approach rather than a fixed rolling window. In this expanding window design, the estimation sample increases by one day with each forecast iteration, starting with a window size of 1000 days (Corsi, 2009; Paye, 2012). This choice is justified by the stability of volatility dynamics captured by HAR models. An expanding window uses all available historical information, thereby improving parameter estimation efficiency under the assumption of parameter stability. Nevertheless, to confirm robustness and explore potential structural changes, a fixed-size rolling-window procedure is also conducted with a window length set to 250, 500, and 1000 observations, as a secondary check.

For each day in the out-of-sample period, 1-day-ahead forecasts of realised variance are computed from each of the competing models (Base-HAR, HAR, HAR-X, HAR-X-LASSO, HAR-X-PCA, and HAR-X-Jump). The forecasts ($\hat{RV}_t$) generated from these models are subsequently compared to the actual observed realised variance (RV_t).

Forecast accuracy is assessed using three distinct error metrics:

1. Mean Squared Error (MSE), defined as:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(\hat{RV}_t - RV_t)^2, \tag{4.9}$$

   which measures average squared deviation and is widely used for general accuracy comparison.

2. The Quasi-likelihood (QLIKE) loss function, calculated as:

$$\text{QLIKE} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{RV_t}{\hat{RV}_t} - \ln\frac{RV_t}{\hat{RV}_t} - 1\right), \tag{4.10}$$

   which is particularly relevant in volatility forecasting due to its scale-invariant property and greater penalty for volatility underprediction, aligning closely with practical risk management objectives (Patton, 2011).

3. Following Paye (2012), we report out-of-sample R^2 statistics for models performing better than the benchmark HAR model, quantifying relative improvements in predictive accuracy. The out-of-sample coefficient of determination is defined as:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{t=1}^{T}(RV_t - \hat{RV}_t)^2}{\sum_{t=1}^{T}(RV_t - \bar{RV})^2}, \tag{4.11}$$

   where $\hat{RV}_t$ denotes the model's predicted value of realised variance at time t, and $\bar{RV}$ is the average realised variance over the evaluation period.

4. To statistically assess whether observed differences in forecast accuracy are significant, we apply the Clark and West (2007) forecast-encompassing test. The test statistic is defined as:

$$CW = \frac{\bar{f}}{\sqrt{\widehat{\text{Var}}(\bar{f})}}, \tag{4.12}$$

   where $\bar{f}$ denotes the sample mean of the adjusted forecast error differentials. The test accounts for the tendency of nested models to overfit and evaluates whether the extended model provides a statistically significant improvement in out-of-sample accuracy.

These accuracy metrics are summarised and compared across models enabling a straightforward assessment of relative forecasting performance.

4.3 Modelling Schematic

We begin by assembling a unified daily panel spanning 2006–2020, merging S&P 500 index data, VIX, realised variance (RV), and diverse investor, consumer, and institutional sentiment and uncertainty indicators - forward-filling lower-frequency (weekly/monthly) series and removing weekends and national holidays for consistent frequency. After log-transformations of RV and VIX, calculating daily returns, and seasonal differencing of nonstationary variables (1-day lag for daily spreads; 22-day lag for surveys and policy measures), stationarity is confirmed via Augmented Dickey-Fuller (ADF) tests.

HAR-style predictors (today's, yesterday's, and 5-, 10-, and 22-day rolling averages of $\log RV$), a binary jump indicator (top 5% threshold of historical $\log RV$), and relevant lags/leads are constructed, with sentiment variables standardised (mean = 0, variance = 1). Four core specifications (and all other models) are estimated using OLS and Newey-West robust standard errors:

1. Base-HAR model.

2. HAR with controls (daily returns, jump dummy, VIX).

3. HAR-X with the full sentiment block.

4. HAR-X-LASSO for penalised variable selection.

Additionally:

- A threshold-HAR-X model explores regime-dependent sentiment impacts based on realised-volatility regimes, with Wald tests for parameter stability.

- Principal Component Analysis (PCA) extracts latent sentiment factors, informing a factor-augmented HAR.

- Jump predictability is examined via a binary-probit HAR-X specification.

Robustness is assessed through rolling-window and subsample analyses, testing forecast stability and parameter sensitivity. Comprehensive diagnostics - including multicollinearity (VIF), residual autocorrelation (Ljung-Box), heteroskedasticity (Breusch-Pagan), and residual normality (Jarque-Bera) - validate the model specifications, complemented by out-of-sample performance metrics (MSE, QLIKE, R^2_{OOS}) and statistical comparison via Clark-West tests.

5 Empirical Results

5.1 In-Sample Estimation Results

The following section presents the empirical results following the methodology presented in Section 4. The results include the four primary regressions performed: Base-HAR, HAR, HAR-X, and HAR-X-LASSO. The results are presented in Table 5.1 and discussed below.

Table 5.1: Primary Results – HAR vs HAR-X vs HAR-X-LASSO Regressions

	Dependent variable:			
	Log Realised Variance			
	Base-HAR	HAR	HAR-X	HAR-X-LASSO
	(1)	(2)	(3)	(4)
logRV	0.449***	0.315***	0.277***	0.277***
	(0.026)	(0.026)	(0.025)	(0.025)
logRV_lag	0.137***	0.113***	0.098***	0.098***
	(0.024)	(0.023)	(0.023)	(0.023)
logRV_lag_week	0.227***	0.192***	0.170***	0.172***
	(0.054)	(0.048)	(0.049)	(0.049)
logRV_lag_biweek	−0.012	−0.042	−0.080	−0.071
	(0.061)	(0.054)	(0.055)	(0.054)
logRV_lag_month	0.151***	−0.030	−0.062	−0.064
	(0.041)	(0.038)	(0.042)	(0.041)
jump		−0.033	−0.045	−0.036
		(0.058)	(0.061)	(0.060)
daily_return		−6.902***	−4.098***	−4.267***
		(1.115)	(1.078)	(1.078)
logVIX		1.196***	1.386***	1.353***
		(0.098)	(0.112)	(0.115)
INV_SKEW			−0.077***	−0.069***
			(0.015)	(0.013)
INV_PCSPX			0.007	
			(0.010)	
INV_BPI			−0.080***	−0.083***
			(0.018)	(0.016)
INV_BULL			−0.008	
			(0.018)	
INV_BEAR			−0.037**	−0.027*
			(0.017)	(0.015)
INV_EXPOSURE			−0.029	−0.017
			(0.019)	(0.018)
INV_EXPOSURE_STD			−0.015	
			(0.011)	

Continued on next page

Table 5.1 – continued from previous page

	Base-HAR (1)	HAR (2)	HAR-X (3)	HAR-X-LASSO (4)
INV_MARGINDEBT			−0.020	−0.020
			(0.013)	(0.013)
CON_SENT			0.009	
			(0.013)	
CON_NEW_BUSINESS			0.026**	
			(0.013)	
CON_REMORTGAGE			−0.001	
			(0.014)	
INST_HYBS			0.040***	0.040***
			(0.012)	(0.012)
INST_TED			−0.035**	−0.034**
			(0.017)	(0.016)
INST_POLICY_UNCERTAINTY			−0.030	−0.028*
			(0.019)	(0.016)
INST_GEOPOL_RISK			0.008	
			(0.040)	
INST_MONETARY_UNCERTAINTY			−0.003	
			(0.020)	
INST_NEWS_SENT			0.007	
			(0.011)	
INST_PRES_APP			−0.022	−0.024
			(0.023)	(0.023)
Constant	−0.472***	−7.971***	−9.932***	−9.732***
	(0.098)	(0.610)	(0.652)	(0.657)
Observations	3,511	3,511	3,511	3,511
R^2	0.750	0.775	0.783	0.783
Adjusted R^2	0.750	0.775	0.782	0.782
Note:				*p<0.1; **p<0.05; ***p<0.01 Robust standard errors computed via Newey–West

Table 5.1 reveals several salient patterns. Across all four specifications, the daily and weekly realised-variance terms remain precisely estimated but their point estimates fall from 0.449/0.227 in Base-HAR to 0.277/0.170 after sentiment is added, indicating that a sizable share of short-horizon persistence is picked up by the exogenous block. The 22-day component is positive and highly significant in the baseline (0.151), yet turns negative (−0.062) once broader information is introduced. The two-day lag is consistently positive with high significance, while the bi-weekly lag remains statistically negligible in every model.

Control variables behave uniformly: $log(VIX)$ retains a stable coefficient near 1.353 to 1.386 in both sentiment regressions, daily absolute returns shrink from −6.902 to −4.098 once sentiment is included, but stay strongly significant, and the jump dummy is never significant. Of the 18 sentiment predictors entered, six achieve at least 10% significance in HAR-X. LASSO retains mostly the same direction and significance for every variable it keeps, however some deviations can be observed. For example, `INST_POLICY_UNCERTAINTY` is not significant in the HAR-X model but is included as a significant variable in the HAR-X-LASSO. Furthermore, `CON_NEW_BUSINESS` is significant in the HAR-X model but is not included in the HAR-X-LASSO. Three predictors - `INV_SKEW` (−0.077), `INV_BPI` (−0.080) and `INST_HYBS` (0.040) - exhibit significant values below the 1% significance level in both models, while `INV_BEAR`, `CON_NEW_BUSINESS` and `INST_TED` show moderate significance. The net effect is an increase in explanatory power from $R^2 = 0.775$ in HAR to 0.783 in HAR-X, with no loss after shrinkage. Adjusted R^2 mirrors this progression, confirming that the additional variables improve fit rather than over-parameterise the equation.

5.2 Additional Results

In addition to the primary results, the in-sample analysis includes secondary results which show three different regressions: HAR-X-Threshold (with the two threshold regressions based on high or low variance), Jump-HAR-X, and HAR-PCA. The results are presented in Table 5.2 and discussed below.

Table 5.2: Combined Secondary Results

	HAR-X-Threshold		Jump-HAR-X	HAR-PCA
	Low Var.%	High Var.%		
	(1)	(2)	(3)	(4)
logRV	0.205***	0.205***		0.304***
	(0.023)	(0.023)		(0.027)
logRV_lag	0.068***	0.068***		0.104***
	(0.020)	(0.020)		(0.022)
logRV_lag_week	0.139***	0.139***		0.178***
	(0.044)	(0.044)		(0.049)
logRV_lag_biweek	−0.080	−0.080		−0.068
	(0.056)	(0.056)		(0.056)
logRV_lag_month	−0.055	−0.055		−0.046
	(0.049)	(0.049)		(0.040)
jump	0.022	0.022	0.155***	−0.108*
	(0.063)	(0.063)	(0.050)	(0.060)
jump_lag			0.145**	
			(0.057)	
jump_lag_week			0.321**	
			(0.144)	
jump_lag_biweek			−0.034	
			(0.146)	
jump_lag_month			0.089	
			(0.093)	
daily_return	−3.345***	−3.345***	−0.681	−5.672***
	(0.985)	(0.985)	(0.540)	(1.264)
logVIX	1.330***	1.330***	0.058***	1.235***

Continued on next page

Table 5.2 – continued from previous page

	Low Var.% (1)	High Var.% (2)	Jump-HAR-X (3)	HAR-PCA (4)
	(0.136)	(0.136)	(0.022)	(0.115)
INV_SKEW	−0.126***	0.099***	0.010***	
	(0.021)	(0.027)	(0.003)	
INV_PCSPX	−0.004	0.019	−0.0003	
	(0.010)	(0.021)	(0.002)	
INV_BPI	−0.130***	0.025	0.001	
	(0.038)	(0.025)	(0.005)	
INV_BULL	0.052**	−0.140***	−0.002	
	(0.026)	(0.032)	(0.005)	
INV_BEAR	−0.108***	0.029	0.003	
	(0.024)	(0.031)	(0.004)	
INV_EXPOSURE	−0.075**	0.046	−0.004	
	(0.036)	(0.032)	(0.007)	
INV_EXPOSURE_STD	−0.026	0.030	−0.003	
	(0.017)	(0.023)	(0.004)	
INV_MARGINDEBT	0.028	−0.019	−0.004	
	(0.024)	(0.021)	(0.005)	
CON_SENT	−0.045**	0.061**	0.005	
	(0.021)	(0.025)	(0.004)	
CON_NEW_BUSINESS	0.060***	−0.016	0.005	
	(0.018)	(0.025)	(0.003)	
CON_REMORTGAGE	0.009	0.021	−0.0004	
	(0.025)	(0.019)	(0.005)	
INST_HYBS	0.067***	0.032**	0.019***	
	(0.020)	(0.013)	(0.005)	
INST_TED	−0.089**	−0.001	0.014**	
	(0.041)	(0.024)	(0.007)	
INST_POLICY_UNCERTAINTY	−0.093***	0.121***	0.003	
	(0.026)	(0.037)	(0.005)	
INST_GEOPOL_RISK	0.254***	−0.534***	−0.013*	
	(0.054)	(0.090)	(0.007)	
INST_MONETARY_UNCERTAINTY	−0.004	−0.073**	0.010	
	(0.037)	(0.033)	(0.006)	
INST_NEWS_SENT	0.002	0.017	−0.001	
	(0.012)	(0.017)	(0.003)	
INST_PRES_APP	0.257***	−0.218***	−0.006	
	(0.052)	(0.057)	(0.007)	
PC1				−0.049***
				(0.011)
PC2				0.023**
				(0.010)
PC3				0.022**
				(0.010)
PC4				0.012
				(0.009)
PC5				0.019

Continued on next page

Table 5.2 – continued from previous page

	Low Var.% (1)	High Var.% (2)	Jump-HAR-X (3)	HAR-PCA (4)
				(0.012)
Constant	−11.012***	−11.012***	−0.157**	−8.836***
	(0.779)	(0.779)	(0.064)	(0.694)
Observations	3,511	3,511	3,511	3,511
R^2	0.829	0.829	0.517	0.779
Adjusted R^2	0.827	0.827	0.513	0.778

Note: HAR-X-Threshold is divided into two columns, but one regression.
Hence, base coefficients and stats are duplicated.
Robust SEs via Newey–West. *p<0.1; **p<0.05; ***p<0.01.

5.2.1 Regime Analysis

Estimating the HAR-X model with sentiment terms interacted with estimated breakpoints lift explanatory power from an adjusted R^2 of 0.782 in the pooled specification to 0.827, indicating that regime heterogeneity captures variation otherwise absorbed by the autoregressive structure.

Investor-oriented proxies exhibit the clearest state dependence. Option skewness (`INV_SKEW`) suppresses volatility in calm markets (−0.126) but amplifies it under stress (+0.099). Statistically significant survey balances reveal that bullish sentiment is volatility-reducing in both low- and high-volatility periods (`INV_BPI`: −0.130 in low volatility and `INV_BULL`: −0.140 in high volatility). Confusingly, bearish sentiment matters only when markets are tranquil, and is also volatility-dampening. Exposure metrics behave similarly where `INV_EXPOSURE` is significantly negative in the low-volatility state, but its effect fade and reverse once the threshold is crossed.

Consumer indicators are asymmetric but weaker. Aggregate sentiment lowers volatility when conditions are calm (−0.045) and increases volatility during periods of turmoil (0.061). New businesses is significant only in the low-volatility regime, and mortgage sentiment is never predictive.

By contrast, institutional uncertainty variables show the largest reversals. Economic policy uncertainty and geopolitical risk shift from −0.093 and +0.254 in the low regime to +0.121 and −0.534 in the high regime, both times highly significant. Presidential approval likewise changes signs, reducing volatility in high-stress periods and increasing it in normal times.

Wald tests (Table A.3) formally corroborate these asymmetries: 14 of the 18 sentiment variables reject the null of equal low- and high-volatility betas at the 5% level, with institutional measures producing the largest statistics.

Collectively, the results demonstrate that sentiment–volatilities are highly regime-specific: variables that dampen realised variance in tranquil markets either weaken or invert during turbulent episodes.

5.2.2 Extreme Event Analysis

The Jump-HAR regression (Table 5.2) shows that the occurrence of past jumps - rather than sentiment - dominates the prediction of next-day jump intensity. The contemporaneous jump indicator and its one-day and one-week lags are all positive and statistically significant (β = 0.155; $p < 0.01$) and (β = 0.145, 0.321; $p < 0.05$), while the bi-weekly lag is insignificant, implying that jump memory decays rapidly beyond one week. Consistent with options-implied risk, $log(VIX)$ remains strongly positive (β = 0.058; $p < 0.01$). Once these jump dummies are included, the sentiment block is slightly washed out: only four variables survive conventional thresholds - `INV_SKEW` (+), `INST_HYBS` (+) and `INST_TED` (+) - alongside a weak negative effect for geopolitical risk. The model's fit (adjusted $R^2 = 0.513$) is well below the 0.782 achieved by the continuous-volatility HAR-X specifications, underscoring that binary jump events are harder to explain and that their predictability is absorbed primarily by their own short-term dynamics rather than by broad sentiment measures.

5.2.3 PCA Analysis

Principal-component analysis on the 18 sentiment and uncertainty series reveals a moderately concentrated factor structure: the first five components account for 59% of total variance (Figure A.5). Loadings (Table A.4) indicate that PC_1 is a broad "optimism" factor - highly positive on investor exposure, option skewness and consumer confidence, and negative on TED-spread and monetary-uncertainty proxies - whereas PC_2 contrasts bullish positioning (negative weights on `INV_BULL` and `INV_BPI`) with consumer optimism (positive weights on `CON_NEW_BUSINESS` and `CON_SENT`). The remaining three components capture narrower themes: policy and investor sentiment (PC_3), dispersion in exposure variability (PC_4), and option-skew versus news tone (PC_5).

When these five orthogonal factors are inserted into the HAR-X framework (Table 5.2), three emerge as statistically relevant. A one-standard-deviation increase in PC_1 lowers next-day realised variance by 4.9% ($p < 0.01$). By contrast, PC_2 raises variance by 2.3% ($p < 0.05$), and PC_3 also carries a positive coefficient (2.2%, $p < 0.05$). PC_4 and PC_5 are insignificant, indicating limited incremental information once the first three factors are controlled for.

The reduced-form model, despite compressing 18 collinear predictors into five uncorrelated factors, attains an $R^2 = 0.779$, and an adjusted $R^2 = 0.778$, only marginally below the full pooled HAR-X specification. Hence, the principal components preserve most of the predictive content of the sentiment block while mitigating degrees-of-freedom and multicollinearity concerns, and they isolate a small subset of economically interpretable latent drivers that materially influence short-horizon volatility.

5.3 Out-of-Sample Forecast Performance Results

In order to test whether the models have predictive power over future volatility, expanding-window out-of-sample tests are performed. The out-of-sample results include R^2_{OOS}, Mean Squared Error (MSE), QLIKE loss function, as well as Clark & West (2007) test. The out-of-sample results are presented in Table 5.3 and discussed below.

Table 5.3: Out-of-Sample Model Evaluation Results

	R^2	MSE	QLIKE	CW t (Base)	CW p (Base)	CW t (HAR)	CW p (HAR)	CW t (LASSO)	CW p (LASSO)
Base-HAR	0.740	0.405	0.260	—	—	—	—	—	—
HAR	0.766	0.374	0.214	12.813	***	—	—	—	—
HAR-X	0.768	0.372	0.200	13.503	***	5.970	***	-0.040	0.516
HAR-X-LASSO	0.770	0.367	0.202	13.818	***	6.639	***	—	—

Note: Clark and West (2007) test statistics are reported for each model against its nested alternative. P-values smaller than 0.001 are marked with *** for simplicity.

Table 5.3 evaluates one-step-ahead forecasts over the out-of-sample period. Relative to the Base-HAR benchmark, successive model augmentations yield monotonic gains. Incorporating control variables raises R^2_{OOS} from 0.740 to 0.766 and lowers MSE and QLIKE by 8% and 18%, respectively. Appending the full sentiment block (HAR-X) delivers an additional, though smaller, improvement (R^2_{OOS} = 0.768; MSE = 0.372; QLIKE = 0.200). Substituting a LASSO-filtered subset of sentiment variables produces the numerically best scores (R^2_{OOS} = 0.770; MSE = 0.367; QLIKE = 0.202), except for QLIKE where HAR-X is slightly lower. This indicates that penalised variable selection refines but does not fundamentally alter predictive accuracy.

Clark-West statistics corroborate these findings. Each enriched specification significantly outperforms Base-HAR at the 1% level ($t \approx 13$). Both sentiment-augmented models also dominate the plain HAR ($t = 5.970$ for HAR-X; $t = 6.639$ for HAR-X-LASSO). The comparison between HAR-X and HAR-X-LASSO is statistically neutral ($t = -0.040$, $p = 0.516$), implying that shrinkage primarily enhances parsimony rather than forecast quality. Collectively, the evidence confirms that sentiment variables provide economically and statistically meaningful gains in volatility prediction beyond the canonical HAR framework.

6 Diagnostics and Robustness Checks

6.1 In-Sample Diagnostic Tests

To evaluate the reliability of the in-sample results, several diagnostics and robustness checks are performed. These tests include tests for autocorrelation (Ljung-Box Q-statistic and residual autocorrelation plots), tests for heteroskedasticity (Breusch-Pagan method), and tests for normality (Shapiro-Wilk and Jarque-Bera). Furthermore, a Q-Q plot of $log(RV)$ is shown to visually assess distributional fit, and the temporal stability of the estimated parameters is shown by expanding window coefficient plots. The results are presented in the following sections.

6.1.1 Autocorrelation

To test for remaining autocorrelation in the residuals, Ljung-Box tests and correlograms are presented.

Table 6.1: Ljung-Box Test Results (lag = 22)

Model	Q-stat	df	p-value
Base-HAR	26.679	22	$***$
HAR	83.094	22	$***$
HAR-X	68.007	22	$***$
HAR-X-LASSO	65.392	22	$***$

Note: P-values smaller than 0.001 are marked with *** for simplicity.

Ljung-Box statistics computed on a 22-day horizon (Table 6.1) uniformly reject the null of serially independent residuals ($p < 0.001$), implying that none of the specifications fully capture the monthly dependence structure of realised variance. The magnitude of the statistic, however, is informative. Base-HAR records the lowest Q-value (26.679), whereas the addition of controls more than doubles the statistic in the plain HAR model (83.094), indicating that the richer model specification leaves more systematic correlation unaccounted for. Incorporating sentiment attenuates this deficit: HAR-X and HAR-X-LASSO reduce the Q-statistic to between 65 and 86, a $\approx$ 20% decline relative to HAR, suggesting that sentiment absorbs part, but not all, of the residual persistence introduced by the control terms.

The correlograms in Figure A.6 corroborate these rankings. Base-HAR shows no bars breaching the 95% confidence envelope. In contrast, the HAR residuals exhibit a pronounced positive autocorrelation at lag 4 followed by a staircase of significant spikes, mirroring the elevated Ljung-Box value. Sentiment-augmented specifications dampen these peaks: HAR-X and its LASSO variant retain significance in the first few lags (except for lag 4) but with lower amplitudes and fewer exceedances beyond lag 8. In sum, while the introduction of sentiment improves in-sample fit and partly mitigates residual dependence, meaningful autocorrelation remains, highlighting the scope for further gains from explicit long-memory or state-space volatility dynamics.

6.1.2 Heteroskedasticity

Table 6.2: Breusch-Pagan Heteroskedasticity Tests

Model	BP Stat	df	p-value
Base-HAR	25.001	5	***
HAR	17.310	8	***
HAR-X	92.693	26	***
HAR-X-LASSO	49.396	20	***

Note: P-values smaller than 0.001 are marked with *** for simplicity.

The OLS implementation of the HAR models assumes that the regression residuals have constant conditional variance. The Breusch-Pagan (BP) test evaluates this assumption by regressing the squared residuals on the fitted values and testing whether the resulting R^2 is zero. A significant statistic therefore signals heteroskedasticity of the error term, i.e., that the dispersion of realised variance around its conditional mean still varies systematically with the state of the market captured by the regressors. In the context of volatility modelling this is not paradoxical: even after accounting for the heterogeneous-autoregressive structure of RV_t, the unexplained component can itself be more volatile when, for instance, recent volatility is high or sentiment is producing a "volatility-of-volatility" effect.

Our results (Table 6.2) illustrate this point. The plain HAR, which already includes multiple lag horizons, yields the smallest BP statistic (17.310), implying relatively mild residual variance clustering. The Base-HAR, lacking controls, performs slightly worse (25.001). Introducing sentiment variables markedly increases the BP statistic (HAR-X: 92.693; HAR-X-LASSO: 49.396), indicating that while these predictors improve mean forecasts they leave larger, structured variance in the errors. The implication is that conventional OLS standard errors would be downward-biased without correction, hence our reliance on Newey-West estimates.

6.1.3 Normality

To assess whether the residuals follow a normal distribution, a Shapiro-Wilk test, Jarque-Bera test, and a visual Q-Q plot of $log(RV)$ are presented. The tests are shown in Table 6.3 and discussed below.

Table 6.3: Residual Normality Test Results

Model	Shapiro-Wilk Stat	p-value	Jarque-Bera Stat	p-value
Base-HAR	0.991	***	199.465	***
HAR	0.989	***	190.023	***
HAR-X	0.989	***	214.074	***
HAR-X-LASSO	0.989	***	200.450	***

Note: P-values smaller than 0.001 are marked with *** for simplicity.

Shapiro-Wilk and Jarque-Bera diagnostics (Table 6.3) reject the null of normally distributed residuals for every specification. The Shapiro-Wilk statistics cluster tightly around 0.99, indicating only modest departures from normality in the central part of the distribution; significance is driven primarily by the large sample size. Jarque-Bera values, however, range from 190 (HAR) to 214 (HAR-X), well above the 5% critical value of 5.99, revealing economically relevant skew-kurtosis combinations. Sentiment-enriched models exhibit marginally larger JB statistics than the baseline specifications, suggesting that adding exogenous predictors falters tail behaviour in the regression errors.

The Q-Q plot of log RV_t (Figure A.7) corroborates these findings: quantiles align closely with the normal benchmark near the median yet diverge in both tails, evidencing excess kurtosis typical of financial volatility even after log transformation. While non-normal errors do not bias OLS coefficients, they invalidate classical t-tests. Consequently, all inference is based on heteroskedasticity- and autocorrelation-consistent (Newey-West) covariances.

6.2 Functional Form and Stability

6.2.1 Alternative Volatility Measures

To assess whether our findings are based on the use of realised variance, we re-estimate the HAR-X specification with the CBOE VIX as the dependent variable. The model achieves an $R^2 = 0.964$ (Table A.5), indicating that the heterogeneous-lag structure coupled with sentiment variables explains almost all day-ahead variation in implied volatility - a higher fit than the 0.783 achieved for realised variance. The own-lag of VIX remains the dominant driver ($\beta_t = 0.897$), while sentiment coefficients exhibit both continuity and divergence relative to the realised-variance results. Only a few predictors retain their earlier signs and significance - where investor sentiment (`INV_BPI`) and (`INV_BEAR`) continue to dampen volatility. Other than that, no other variable that is significant in the VIX regression is significant in the HAR-X model in the primary results.

Overall, the high explanatory power and the partial alignment of key coefficients corroborate the robustness of our main conclusions: sentiment carries incremental predictive content even when volatility is measured from option prices rather than high-frequency returns. Differences in significance across specifications underscore that realised and implied volatility capture distinct facets of market risk, ex-post variation versus ex-ante expectations, so that sentiment can load differently on each.

6.3 Sample Robustness

To further evaluate the robustness of our results, this section evaluates a subsample analysis of all the results, as well as analyses the coefficient stability. These checks provide further confidence for whether the presented results remain robust under alternative model setups.

6.3.1 Subsample Analysis

To measure temporal stability, each HAR-family specification is re-estimated on the first and second halves of the sample (Table A.6) meaning a breakpoint in 2013. As expected, goodness-of-fit metrics decline when the estimation window is halved, yet the ranking of models is preserved: HAR-X and HAR-X-LASSO continue to dominate with R^2 values near 0.75 in both subperiods, whereas the Base-HAR remains the weakest. Mean-squared error is remarkably stable across splits, indicating that forecast accuracy does not deteriorate materially when the models are trained on fewer observations.

Coefficient patterns are mostly invariant, but several notable sign reversals emerge. Investor pessimism (`INV_BEAR`) is volatility-reducing in the early subsample and volatility-enhancing later, and investor optimism (`INV_BULL`) exhibits a similar image - consistent with a shift from contrarian to pro-cyclical sentiment after the financial crisis. The `INV_BULL` also loses significance in the full sample despite being significant in both halves, implying that its effect is driven by episodic spikes rather than a persistent relationship. Consumer confidence (`CON_SENT`) is insignificant and positive pre-2013, yet turns negative and significant thereafter, suggesting that household sentiment became a more reliable barometer of risk in the low-rate, post-QE environment.

Conversely, option skew (`INV_SKEW`) and broker positioning breadth (`INV_BPI`) remain negative and highly significant in every split and specification, underscoring their robustness as forward indicators of realised variance. The mixed survivorship of other predictors, e.g., `CON_NEW_BUSINESS` and `INST_HYBS` mattering in only one subperiod, suggests that their explanatory power is regime-dependent. Overall, the subsample exercise confirms that the core findings are not artefacts of a particular market phase while highlighting that certain sentiment channels evolve with underlying macro-financial conditions.

6.3.2 Coefficient Stability

To trace how the link between sentiment and volatility evolves through time, we re-estimate the HAR-X model on an expanding window and plot the cumulative coefficient paths for each sentiment regressor (Figure A.9). Two clear patterns emerge. First, a core set of predictors display stability: option skew (`INV_SKEW`), put–call demand (`INV_PCSPX`), institutional high-yield spread (`INST_HYBS`), broker positioning breadth (`INV_BPI`), and new-business applications (`CON_NEW_BUSINESS`) maintain both sign and magnitude as the sample grows, supporting their interpretation as structurally persistent drivers of realised variance. Second, a subset of variables undergoes sign reversals, signalling regime-dependent effects. Retail optimism (`INV_BULL`) and dispersion in institutional exposure (`INV_EXPOSURE_STD`) flip from volatility-reducing in the post-crisis period to volatility-enhancing during the low-rate era, while household sentiment (`CON_SENT`) and refinancing activity (`CON_REMORTGAGE`) turn from weakly positive to markedly negative after 2013. Inversions are observed for monetary policy and TED-spread measures, indicating that their impact on volatility depends on the prevailing policy regime and liquidity backdrop.

Overall, the plots reinforce earlier sub-sample evidence: a stable "spine" of sentiment variables contributes consistently to volatility forecasts, whereas others behave cyclically, amplifying or dampening volatility according to macro-financial conditions.

6.4 Forecasting Metric Robustness

To ensure that our out-of-sample tests are not sensitive to a specific window length, a rolling-window evaluation is performed of the out-of-sample tests (R^2_{OOS}, Mean Squared Error (MSE), QLIKE loss function, Clark & West (2007)) with 250-day, 500-day and 1000-day windows. The result is presented in Table 6.4 with results discussed below.

Table 6.4: Out-of-Sample Rolling-Window Evaluation

Window	Model	OOS R^2	MSE	QLIKE	CW t-stat	CW p-val
1000-day	Base-HAR	0.677	0.424	0.265	NA	—
1000-day	HAR	0.715	0.375	0.226	13.381	***
1000-day	HAR-X	0.715	0.375	0.230	14.036	***
1000-day	HAR-X-LASSO	0.717	0.372	0.228	14.142	***
250-day	Base-HAR	0.659	0.403	0.253	—	—
250-day	HAR	0.693	0.363	0.219	14.926	***
250-day	HAR-X	0.662	0.399	0.243	14.517	***
250-day	HAR-X-LASSO	0.682	0.376	0.225	14.880	***
500-day	Base-HAR	0.711	0.405	0.254	—	—
500-day	HAR	0.741	0.363	0.218	14.298	***
500-day	HAR-X	0.738	0.367	0.223	14.883	***
500-day	HAR-X-LASSO	0.741	0.364	0.219	14.746	***

Note: P-values smaller than 0.001 are marked with *** for simplicity.

As seen in the table above, the rolling-window estimation shows a relatively nuanced result, with no model consistently having the best or worst results. Looking at R^2_{OOS}, the models perform very varying during the different time horizons, with HAR-X-LASSO having the highest estimate at a 1000-day horizon and HAR the highest at a 250-day horizon and both having the same number at a 500-day horizon. HAR-X has a higher estimate for longer horizons but drops to a lower value in the 250-day estimate. The MSE and QLIKE tests confirm this result by having very similar relationships between the models. In QLIKE, the HAR model has consistently the lowest value across all time horizons, and the HAR-X-LASSO consistently ranks second. In both the MSE and the QLIKE, HAR-X performs better at a longer horizon, suggesting that model complexity can backfire in very short time horizons. The Clark-West test shows that all models, even HAR-X, significantly outperform the Base-HAR model across all time horizons with p-values lower than 0.001 confirming that the gains in forecasting are statistically significant. However, even if sentiment-enhanced models improve explanatory power, the results show that they do not always outperform simpler models in terms of penalised forecast error, especially in a shorter window.

7 Discussion

The question that this paper aims to answer is whether sentiment variables hold significant prediction power on realised market volatility. Furthermore, the paper explores different types of sentiment (investor, public, and institutional) to analyse whether there is consistency within the groups and if any group performs better than the others, as well as an analysis of regime dependency.

7.1 Predictive Value of Sentiment-Enhanced Models

Empirical analysis concludes that the HAR-X model and the HAR-X-LASSO model outperform the more simple models like the Base-HAR and HAR with control variables. A.8 shows how Base-HAR consistently over-estimates realised variance, in comparison to sentiment-augmented models that reduce cumulative forecast error. This is true for both in- and out-of-sample analysis. These results indicate that sentiment variables have incremental information over what the traditional variables have in forecasting realised variance, which is in line with previous literature (e.g., Verma and Verma, 2007; Kelly and Jiang, 2014; Patton and Sheppard, 2015; Mittnik et al, 2015; Audrino et al, 2020; Ding et al, 2021; Bai et al, 2024; Li et al, 2024). The improvement can be seen by the increase in R^2 for both the HAR-X model and the HAR-X-LASSO, as well as the reduction in out-of-sample forecast error metrics such as MSE and QLIKE. The result, thereby contributes to the aforementioned papers in the context of the HAR framework by showing that sentiment variables are significant predictors even after accounting for autoregressive features of volatility through HAR lags, as well as well-established benchmark variables such as VIX. However, it is important to note that the magnitude of the improvements is relatively small, with out-of-sample tests for HAR-X and HAR-X-LASSO showing improvements of no more than 7-10% compared to the base models. The improvements are statistically significant as shown by the Clark & West test, but imply that the practical and economic improvements are limited. Overall, the evidence supports the hypothesis that sentiment contains information that could be valuable in forecasting stock market volatility.

7.2 Relevance of Different Sentiment Categories

Comparing the three groups - investor, public, and investor sentiment - is useful to understand the underlying drivers for these sentiment variables in predicting future volatility. By identifying which type of sentiment has the highest incremental power, the understanding of the relationship between sentiment and volatility becomes more nuanced and may contribute to a deeper understanding.

It can be concluded that investor sentiment has the most consistency in predictability for realised variance. The variables included in this sentiment group generally act as theory would suggest (e.g. Lee et al, 2002; Baker and Wurgler, 2006; Mora-Valencia et al, 2021), with SKEW and BPI receiving constantly significant results. Furthermore, LASSO consistently favours many of these variables, suggesting that they give incremental explanations for forecasting volatility beyond the fundamentals.

The public sentiment group gives mixed results, with some variables being insignificant, or showing coefficient signs that were inconsistent with theory (see Section 7.4 for more in-depth analysis). Interestingly, the influence of the predictors seems to be dependent on condition, both regarding period, and volatility. For example, the Michigan Consumer Sentiment Index was only significant in the latter half of the subsample regression, showing how the variable is affected by the economic environment. This is also true for the number of new business applications, which are only significant for periods of lower volatility. This state-dependency shows that public sentiment variables are informative only in certain cases and that their impact reflects longer-term market trends rather than short horizons.

The institutional sentiment variables show a more complex behaviour compared to the other groups, having clear dependence on the market regime, but do not behave uniformly. The only truly stable variable is high-yield bond saving, which, similarly to the results by Chun et al (2023), shows strong incremental power. The economic policy uncertainty index behaves as theory describes, showing significant and positive results during turbulent times while its predictive power has a muted effect during calmer periods. This aligns with Pástor and Veronesi (2013) who argue that uncertainty has a stronger effect on risk premia during periods of stress. Furthermore, the geopolitical risk indicator, while insignificant in the main model, contradicts previous literature (Caldara and Iacoviello, 2013) by being negatively significant during turbulent times, which would indicate that higher geopolitical risks lower volatility. However, looking at the subsample analysis, GPR is not significant in either of the subsamples, indicating some instability with the variable. Monetary policy uncertainty indicates similar behaviours, being insignificant in the main model, but negative and significant during turbulent times. While inconsistent with theory, it aligns with the findings by Hsiao et al (2022). Overall, the variables show the complexity of finding truly powerful predictors of volatility.

In general, our results suggest that volatility forecasting models favour market-driven and trading-oriented variables, while broader macroeconomic sentiment requires further analysis.

7.3 Robustness and Model Behavior Over Time

To verify the analysis and examine whether the results are sample-specific, several robustness tests are conducted, changing both the sample period and model approach. In general, the results show relatively high robustness, with some deviations.

Looking at model consistency, out-of-sample rolling windows evaluation shows that sentiment-based models generally perform better in longer windows, with the 250-day window favouring the more simple models. However, HAR-X-LASSO shows that mitigating overfitting concerns helps even in the short term. For all horizons, the improvements were all significant, tested through a Clark and West (2007) test. These results are in line with Audrino et al (2020) showing that sentiment has incremental power for out-of-sample results. The result highlights the importance of either including a long estimation window, or accounting for overfitting through regularisation when using many variables as done in this paper.

Furthermore, the results from the subsample test also reflect relatively robust variables, with the exceptions previously mentioned. The gradual change in some of the variables in the subsample could be affected by the two periods that the subsample reflects, pre, and post, 2013, where structural changes in the market can affect the result. This suggests that one should be mindful of the occurrence of nonstationarity when handling variables such as sentiment as they are closely tied to the economic environment and market structure. Threshold analysis, where the variables were tested in high-volatility and low-volatility environments, helps identify breaks in market structure. The results of that reinforce that financial context is important when analysing volatility.

In addition to stability, this thesis also tests for dimensionality to ensure that the results of our analysis are not a result of how we include sentiment. This was done using dimensionality reduction through a Principal Component Analysis (PCA) with significant results for the first three principal components. However, the PCA did not produce higher R^2 than the HAR-X-LASSO showing that even if there is a low dimensionality in the sentiment measures, keeping fewer individual sentiments in the analysis is more advantageous for prediction than a summary of all sentiments. This also implies that there are a few sentiments that explain most of the volatility and are by themselves genuinely informative.

7.4 Interpretation of Sentiment Effects and Sign Deviations

Several variables do not act according to what theory would expect. `INV_BEAR` has a negative significant coefficient in both the HAR-X model and the HAR-X-LASSO. This contradicts the theory which explains that negative or uncertain investor sentiment should increase volatility (Kresta et al, 2024). The reason for this could be many but may involve a lagging relationship where the variable increases only after volatility already has peaked. However, looking at the secondary result, `INV_BEAR` is only significant and negative in the low variance threshold, implying that when volatility is low, increased caution lowers volatility further. Furthermore, `INST_POLICY_UNCERTAINTY` also has a negative significant relationship with volatility, which deviates from the theory saying that increased policy uncertainty should increase volatility (Liu and Zhang, 2015). However, secondary results show that this is only true when volatility is low. When volatility is high, an increase in economic policy uncertainty increases volatility significantly. Moreover, the CON_NEW_BUSINESSES variable is significant in a positive relation with volatility, suggesting that increased business formations increase volatility, again contradicting prior research (e.g. Bonato et al, 2024). However, looking at the secondary results, one can see that this is only true during a period of low volatility. This is also true for INST_TED which has a negative significant relationship with volatility, suggesting that an increased TED-spread lowers volatility, contradicting prior theory (e.g. Mittnik et al, 2015).

Another interesting variable is the INV_SKEW, which has a negative significant coefficient in the HAR-X and HAR-X-LASSO models. This would imply that an increase in SKEW predicts a decrease in volatility, which contradicts prior research (Mora-Valencia et al, 2021). However, looking more closely at the secondary results, it can be observed that SKEW is significantly negatively correlated with volatility for the low variance threshold, and has a positive significant relationship with volatility in the high variance threshold. This implies that when volatility is low, increased SKEW lowers volatility, whereas when volatility is high, increased SKEW increases volatility.

However, when explaining these inconsistencies, methodological or date-related factors must also be considered. One reason for these deviations could be multicollinearity where many of the included variables tend to move together (as shown in Table A.2). This can cause inconsistencies which results in variables not showing their true effects. To mitigate this, LASSO was added as an approach, but it does not fully remove the correlation between the included variables. Furthermore, mismatches in lags may also affect the results. As the included variables had different frequencies in the obtained data, but were still interpolated into our daily model, the models with lower frequency may already have priced in information which may, when changed to a daily level, damper or reverse the result on a daily level. Lastly, attributes may be missed when performing a linear model if there are nonlinear effects in the data. Although the extended analysis included both threshold and subsample analysis to mitigate this, further complexity was not explored. Therefore, the findings encourage a nuanced interpretation rather than an acceptance at face value.

7.5 Practical Implications for Volatility Forecasting

While the paper has a theoretical approach, the results are relevant to practitioners, for example, within trading and risk management. The findings show that incorporating sentiment variables into volatility prediction can improve forecasts, at least on the margin. Even small gains can be valuable for practitioners, who can use variables such as option-implied indices or investor sentiment surveys as signals for a future change in volatility. Although our research supports the inclusion of VIX - a very common measurement among practitioners - broader sentiment variables can capture complementary information that refine the forecasts. Furthermore, the analysis shows that volatility forecasting in practice might be improved by more complex modelling where market conditions are considered, as many variables acted differently when volatility was high and low. Finally, the results have implications for policymakers looking at stability in the financial market, as the result showed deviating results in different uncertainty indices, suggesting a more nuanced approach for practitioners.

7.6 Limitations and Future Research

While the models in this paper are robust, several limitations exist that might affect the outcome. Recognising these limitations both contextualises the findings, as well as suggests possible future research.

One limitation of this paper is data timeliness. Much of the sentiment data used was not available in real time or at a daily frequency. Several proxies (e.g. surveys and uncertainty indices) are, therefore, lagged either weekly or monthly assuming a carrying-forward of the last value. This limits the interpretation of the results as it idealises the forecasting scenario. The results should, hence, be viewed in the context of this, and the resulting forecasts can be interpreted as an upper bound on predictive performance which limits the immediate practical applicability of the models.

Furthermore, incorporating 18 variables with data within a limited area raises the risk of overfitting the model. This was mitigated to an extent using LASSO and robustness checks, but the problem of sample-specific results remains. Out-of-sample tests and subsample analysis were included to ensure genuine patterns, but when using this many variables it cannot be completely ensured. Therefore, future research could implement more effective robustness techniques such as random forests or neural networks that can handle this problem more efficiently.

Another limitation of this study is that it is solely focused on the S&P 500, which limits the data to U.S.-specific data. The paper does not analyse whether the results would be different in markets where areas such as political or economic stability differ. For future research, it would be interesting to put this model into the context of other markets to see if the trends observed here are true for the broader financial market.

Moreover, this paper used realised variance as the dependent variable to forecast one day ahead. Although this is the most common approach, there are many ways to measure volatility such as implied volatility or volatility over longer horizons, which could, if analysed, provide further robustness to our results. Additionally, our data was limited between 2006 and 2020 due to data availability, which limits a true long-term analysis.

Finally, the economic interpretation of our results is limited by the abstractness of some of the included variables. While this paper tries to understand each variable as extensively as possible, there are still gaps behind what fully drives each variable.

Looking ahead, there are several interesting developments of this paper. Firstly, examining more sophisticated sentiment data would offer a more comprehensive understanding of the area, including high-frequency variables or deriving data from alternative sources such as social media and financial news. This would provide conclusions from other areas of sentiment that this paper did not cover. Furthermore, future models could incorporate regime-switching, or machine learning, to capture the nonlinear effects of the variables more comprehensively, which this paper chose to not look at. Another interesting topic to explore is whether the divergence of sentiment affects volatility, meaning that volatility increases when investors or the public disagree on their opinion regarding the economic outlook.

In conclusion, our research provides evidence that including sentiment in volatility forecasting models enhances predictability as well as gives insights into what type of sentiment matters and when. By acknowledging its limitations, future research can further explore the complexities that sentiment has in the broader financial market. Such efforts would not only increase the accuracy but also deepen our understanding of the behavioural forces that shape the market.

8 Conclusion

This thesis investigated whether sentiment-based indicators can enhance the forecasting of stock market volatility through an HAR framework. Building on the structure used by Paye (2012) and extending it to sentiment variables, we aimed to bridge the gap between traditional volatility models and insights from behavioural finance.

Through a comprehensive analysis of the S&P 500 market, this paper concludes that including sentiment variables increases the predictive power of volatility both in an in-sample analysis and in an out-of-sample analysis, compared to a more primitive HAR model. The results show that investor sentiment variables, such as SKEW and BPI, were particularly informative, whereas public sentiment and institutional sentiment revealed a more nuanced picture of the incremental power of sentiment. By including dimensionality reduction and variable selection through LASSO and PCA, as well as several robustness analyses, the results were enhanced, highlighting the robustness of sentiment-based improvements in volatility forecasting.

These findings contribute to the broad scope of literature supporting the role of sentiment in financial markets, suggesting that their incorporation can improve the efficiency of volatility forecasting. This can be used by practitioners working with portfolio allocation, derivative pricing, as well as risk management.

While the results are tested for robustness, certain limitations should be acknowledged, including the lack of real-time availability of data, and the risk of overfitting when including a broad range of variables. Furthermore, the paper is limited to the S&P 500 which limits the generalisation of the result across markets.

Future research can expand on this paper by implementing the same method in other markets to test the robustness of the results from a broader perspective. Moreover, it would be interesting to examine a more sophisticated set of sentiment variables, which can enrich the results, as well as incorporating regime-switching models to further account for nonlinear effects. Further exploration of how divergence of sentiment might affect volatility also presents an interesting avenue for deeper behavioural insights.

Overall, this thesis highlights the relevance of moods and behavioural signals across the financial markets when analysing future risk. By systematically integrating sentiment into volatility forecasting, we show that markets not only have memory - but mood also matters.

References

Literature

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial Intermediaries and the Cross-Section of Asset Returns, *Journal of Finance* 69, 2557–2596. https://doi.org/10.1111/jofi.12189

Andersen, Torben G., and Tim Bollerslev, 1998, Answering the skeptics: Yes, standard volatility models do provide accurate forecasts, *International Economic Review* 39, 885–905. https://doi.org/10.2307/2527343

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, 2003, Modeling and Forecasting Realized Volatility, *Econometrica* 71, 579–625. https://doi.org/10.1111/1468-0262.00418

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold, 2007, Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility, *Review of Economics and Statistics* 89, 701–720. https://doi.org/10.1162/rest.89.4.701

Asgharian, Hossein, Ai J. Hou, and Farrukh Javed, 2013, The importance of the macroeconomic variables in forecasting stock return variance: A GARCH-MIDAS approach, *Journal of Forecasting* 32, 600–612. https://doi.org/10.1002/for.2256

Audrino, Francesco, Fabio Sigrist, and Daniele Ballinari, 2020, The impact of sentiment and attention measures on stock market volatility, *International Journal of Forecasting* 36, 334–357. https://doi.org/10.1016/j.ijforecast.2019.05.010

Bai, Sarula, Jaewon Jung, and Shun Li, 2024, The Spillover Effects of Market Sentiments on Global Stock Market Volatility: A Multi-Country GJR-GARCH-MIDAS Approach, *Journal of Risk and Financial Management* 17, 569. https://doi.org/10.3390/jrfm17120569

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2016, Measuring Economic Policy Uncertainty, *Quarterly Journal of Economics* 131, 1593–1636. https://doi.org/10.1093/qje/qjw024

Baker, Malcolm, and Jeffrey Wurgler, 2006, Investor Sentiment and the Cross-Section of Stock Returns, *Journal of Finance* 61, 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Baker, Malcolm, and Jeffrey Wurgler, 2007, Investor Sentiment in the Stock Market, *Journal of Economic Perspectives* 21, 129–152. https://doi.org/10.1257/jep.21.2.129

Barber, Brad M., and Terrance Odean, 2008, All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors, *Review of Financial Studies* 21, 785–818. https://doi.org/10.1093/rfs/hhm079

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A Model of Investor Sentiment, *Journal of Financial Economics* 49, 307–343. https://doi.org/10.1016/S0304-405X(98)00027-0

Barndorff-Nielsen, Ole E., and Neil Shephard, 2004, Power and bipower variation with stochastic volatility and jumps, *Journal of Financial Econometrics* 2, 1–37. https://doi.org/10.1093/jjfinec/nbh001

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi, 2018, Do ETFs Increase Volatility?, *Journal of Finance* 73, 2471–2535. https://doi.org/10.1111/jofi.12727

Black, Fischer, 1976, Studies of stock price volatility changes, *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association*, 177–181.

Blair, Bevan, Ser-Huang Poon, and Stephen J. Taylor, 2000, Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High Frequency Index Returns, *Lancaster University Management School, Working Paper No. 99/014*. http://dx.doi.org/10.2139/ssrn.182128

Bodilsen, Simon T., and Asger Lunde, 2025, Exploiting News Analytics for Volatility Forecasting, *Journal of Applied Economics* 40, 18–36. https://doi.org/10.1002/jae.3095

Bollerslev, Tim, and Viktor Todorov, 2011, Tails, fears, and risk premia, *The Journal of Finance* 66, 2165–2211. https://doi.org/10.1111/j.1540-6261.2011.01695.x

Bonato, Matteo, Oguzhan Cepni, Rangan Gupta, and Christian Pierdzioch, 2024, Business applications and state-level stock market realized volatility: A forecasting experiment, *Journal of Forecasting* 43, 456–472. https://doi.org/10.1002/for.3042

Brown, Gregory W., 1999, Volatility, Sentiment, and Noise Traders, *Financial Analysts Journal* 55, 82–90. https://doi.org/10.2469/faj.v55.n2.2263

Brown, Gregory W. and Michael T. Cliff, 2004, Investor sentiment and the near-term stock market, *Journal of Empirical Finance* 11, 1-27. https://doi.org/10.1016/j.jempfin.2002.12.001

Caldara, Dario, and Matteo Iacoviello, 2018, Measuring Geopolitical Risk, *International Finance Discussion Papers* 1222. https://doi.org/10.17016/IFDP.2018.1222

Carroll, Christopher D., Jeffrey C. Fuhrer, and David W. Wilcox, 1994, Does Consumer Sentiment Forecast Household Spending? If So, Why?, *American Economic Review* 84, 1397–1408. http://www.jstor.org/stable/2117779

Chen, Sipeng, and Gang Li, 2021, Why Does Option Implied Volatility Forecast Realized Volatility? Evidence from News Events, *Hong Kong Polytechnic University Working Paper*. http://dx.doi.org/10.2139/ssrn.3905677

Christie, Andrew A., 1982, The stochastic behavior of common stock variances: Value, leverage and interest rate effects, *Journal of Financial Economics* 10, 407–432. https://doi.org/10.1016/0304-405X(82)90018-6

Chun, Dohyun, Hoon Cho, and Doojin Ryu, 2023, Discovering the drivers of stock market volatility in a data-rich world, *Journal of International Financial Markets, Institutions and Money* 82, 101684. https://doi.org/10.1016/j.intfin.2022.101684

Clark, Todd E., and Kenneth D. West, 2007, Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* 138, 291–311. https://doi.org/10.1016/j.jeconom.2006.05.023

Corsi, Fulvio, 2009, A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics* 7, 174–196. https://doi.org/10.1093/jjfinec/nbp001

Corsi, Fulvio, and Roberto Renò, 2010, HAR volatility modelling with heterogeneous leverage and jumps, Unpublished paper, Università di Siena.

De Bondt, Werner F. M., and Richard Thaler, 1985, Does the Stock Market Overreact?, *Journal of Finance* 40, 793–805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Noise Trader Risk in Financial Markets, *Journal of Political Economy* 98, 703–738. http://www.jstor.org/stable/2937765

Dickey, David A., and Wayne A. Fuller, 1979, Distribution of the Estimators for Autoregressive Time

Series With a Unit Root, *Journal of the American Statistical Association* 74, 427–431. https://doi.org/10.2307/2286348

Ding, Yi, Dimos Kambouroudis, and David G. McMillan, 2021, Forecasting Realised Volatility: Does the LASSO approach outperform HAR?, *Journal of International Financial Markets, Institutions and Money* 74, 101386. https://doi.org/10.1016/j.intfin.2021.101386

Enders, Walter, 2014, *Applied Econometric Time Series*, 4th ed, John Wiley & Sons.

Engle, Robert F., and Jose Gonzalo Rangel, 2008, The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes, *Review of Financial Studies* 21, 1187–1222. https://doi.org/10.1093/rfs/hhn004

Fama, Eugene F., 1965, The Behavior of Stock Market Prices, *Journal of Business* 38, 34–105. http://www.jstor.org/stable/2350752

Fortune, Peter, 2001, Margin Lending and Stock Market Volatility, *New England Economic Review* 4, 3–25.

Gang, Jianhua, Nan Huang, Ke Song, and Ruyi Zhang, 2020, Index Volatility and the Put–Call Ratio: A Tale of Three Markets, *Quantitative Finance* 20, 1983–1996. https://doi.org/10.1080/14697688.2020.1814009

Garcia, Maria T. and Carolina S. C. Carvalho, 2025, Measuring Sentiment: The Impact on Financial Markets Volatility, *REM Working Paper Series* 0365/2025. http://hdl.handle.net/10400.5/97423

Granger, Clive W. J., and Paul Newbold, 1974, Spurious regressions in econometrics, *Journal of Econometrics* 2, 111–120. https://doi.org/10.1016/0304-4076(74)90034-7

Gupta, Rangan, Yuvana Jaichand, Christian Pierdzioch, and Reneé van Eyden, 2023, Realized Stock-Market Volatility of the United States and the Presidential Approval Rating, *Mathematics* 11, 2964. https://doi.org/10.3390/math11132964

Hamilton, James D., 1994, Time Series Analysis, *Princeton University Press*. https://doi.org/10.2307/j.ctv14jx6sm

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York).

Heber, Gerd, Asger Lunde, Neil Shephard, and Kevin Sheppard, 2009, *Oxford-Man Institute's realized library*, Oxford-Man Institute, University of Oxford.

Hong, Harrison, and Jeremy C. Stein, 2007, Disagreement and the Stock Market, *Journal of Economic Perspectives* 21, 109–128. https://doi.org/10.1257/jep.21.2.109

Hsiao, Shisong, Chaoqun Ma, Ting Zhang, and Liurui Deng, 2022, Monetary Policy Uncertainty and Stock Market Volatility, *SSRN Working Paper*. http://dx.doi.org/10.2139/ssrn.4215281

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2021, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed., Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jena, Sangram Keshari, Aviral Kumar Tiwari, and Amarnath Mitra, 2019, Put–Call Ratio Volume vs. Open Interest in Predicting Market Return: A Frequency Domain Rolling Causality Analysis, *Economies* 7, 24. https://doi.org/10.3390/economies7010024

Kelly, Bryan, and Hao Jiang, 2014, Tail Risk and Asset Prices, *The Review of Financial Studies* 27, 2841–2871. http://www.jstor.org/stable/24466856

Keys, Benjamin J., Tomasz Piskorski, Amit Seru, and Vincent Yao, 2014, Mortgage Rates, Household Balance Sheets, and the Real Economy, *NBER Working Paper* 20561.

http://doi.org/10.3386/w20561

Kresta, Aleš, Jialei Xiong, and Bahate Maidiya, 2024, Sentiment and Stock Characteristics: Comprehensive Study of Individual Investor Influence on Returns, Volatility, and Trading Volumes, *Business Systems Research Journal* 15, 67–82. https://doi.org/10.2478/bsrj-2024-0018

Kyle, Albert S., 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315–1335. https://doi.org/10.2307/1913210

Li, Xiafei, Chao Liang, and Feng Ma, 2022, Forecasting Stock Market Volatility with a Large Number of Predictors: New Evidence from the MS-MIDAS-LASSO Model, *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04716-1

Li, Zhao-Chen, et al., 2024, Forecasting Global Stock Market Volatilities: A Shrinkage Heterogeneous Autoregressive (HAR) Model with a Large Cross-Market Predictor Set, *International Review of Economics & Finance* 93, 673–711. https://doi.org/10.1016/j.iref.2024.05.008

Liu, Li, and Tao Zhang, 2015, Economic Policy Uncertainty and Stock Market Volatility, *Finance Research Letters* 15, 99–105. https://doi.org/10.1016/j.frl.2015.08.009

Lindblad, Annika, 2017, Sentiment Indicators and Macroeconomic Data as Drivers for Low-Frequency Stock Market Volatility, *MPRA Paper* 80266. https://mpra.ub.uni-muenchen.de/id/eprint/80266

Lee, Wayne Y., Christine X. Jiang, and Daniel C. Indro, 2002, Stock Market Volatility, Excess Returns, and the Role of Investor Sentiment, *Journal of Banking & Finance* 26, 2277–2299. https://doi.org/10.1016/S0378-4266(01)00202-3

Mandelbrot, Benoît, 1963, The Variation of Certain Speculative Prices, *Journal of Business* 36, 394–419. https://www.jstor.org/stable/2350970

Mian, Atif, and Amir Sufi, 2011, House Prices, Home Equity-Based Borrowing, and the U.S. Household Leverage Crisis, *American Economic Review* 101, 2132–2156. https://doi.org/10.1257/aer.101.5.2132

Mittnik, Stefan, Nikolay Robinzonov, and Martin Spindler, 2015, Stock Market Volatility: Identifying Major Drivers and the Nature of Their Impact, *Journal of Banking & Finance* 58, 1–14. https://doi.org/10.1016/j.jbankfin.2015.04.003

Mora-Valencia, Andrés, Santiago Rodríguez-Raga, and Esteban Vanegas, 2021, Skew Index: Descriptive Analysis, Predictive Power, and Short-Term Forecast, *North American Journal of Economics and Finance* 56, 101356. https://doi.org/10.1016/j.najef.2020.101356

Newey, Whitney K., and Kenneth D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703–708. https://doi.org/10.2307/1913610

Pan, Jun, and Allen M. Poteshman, 2006, The Information in Option Volume for Future Stock Prices, *Review of Financial Studies* 19, 871–908. https://doi.org/10.1093/rfs/hhj024

Pástor, Ľuboš, and Pietro Veronesi, 2013, Political Uncertainty and Risk Premia, *Journal of Financial Economics* 110, 520–545. https://doi.org/10.1016/j.jfineco.2013.08.007

Patton, Andrew J., 2011, Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* 160, 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Patton, Andrew J., and Kevin Sheppard, 2015, Good volatility, bad volatility: Signed jumps and the persistence of volatility, *The Review of Economics and Statistics* 97, 683–697. https://doi.org/10.1162/REST_a_00503

Paye, Bradley S., 2012, Déjà Vol: Predictive Power of Macro-Finance Variables for Equity Market

Volatility, *Journal of Financial Economics* 106, 527–546.
`https://doi.org/10.1016/j.jfineco.2012.06.005`

Shiller, Robert J., 2003, From Efficient Markets Theory to Behavioral Finance, *Journal of Economic Perspectives* 17, 83–104. `https://doi.org/10.1257/089533003321164967`

Shleifer, Andrei, and Robert W. Vishny, 1997, The Limits of Arbitrage, *Journal of Finance* 52, 35–55.
`https://doi.org/10.1111/j.1540-6261.1997.tb03807.x`

Sum, Vichet, and Jack Chorlian, 2013, Stock Market Risk Premiums, Business Confidence and Consumer Confidence: Dynamic Effects and Variance Decomposition, *International Journal of Economics and Finance* 5, 45–49. `https://doi.org/10.5539/ijef.v5n9p45`

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *Journal of Finance* 62, 1139–1168. `https://doi.org/10.1111/j.1540-6261.2007.01232.x`

Tibshirani, Robert, 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288.
`https://doi.org/10.1111/j.2517-6161.1996.tb02080.x`

Tsay, Ruey S., 2005, *Analysis of Financial Time Series*, 2nd ed, John Wiley & Sons.
`https://doi.org/10.1002/0471746193`

Verma, Rahul, and Priti Verma, 2007, Noise Trading and Stock Market Volatility, *Journal of Multinational Financial Management* 17, 231–243.
`https://doi.org/10.1016/j.mulfin.2006.10.003`

Yu, Jianfeng, and Yu Yuan, 2011, Investor Sentiment and the Mean–Variance Relation, *Journal of Financial Economics* 100, 367–381. `https://doi.org/10.1016/j.jfineco.2010.10.011`

Zivot, Eric, and Jiahui Wang, 2006, *Modeling Financial Time Series with S-PLUS*, 2nd ed, Springer.
`https://doi.org/10.1007/978-0-387-32348-0`

Zou, Hui, 2006, The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association* 101, 1418–1429. `https://doi.org/10.1198/016214506000000735`

Data Sources

American Association of Individual Investors, 2025, AAII Investor Sentiment Survey, (AAII).

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2025, Monetary Policy Uncertainty Index, (Economic Policy Uncertainty).

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2025, US Economic Policy Uncertainty Index, (Economic Policy Uncertainty).

Caldara, Dario, and Matteo Iacoviello, 2021, Geopolitical Risk Index, (Economic Policy Uncertainty).

Cboe Global Markets, 2025, VIX Index Historical Data, (Cboe).

Cboe Global Markets, 2025, SKEW Index Dashboard, (Cboe).

CRAN, 2024, rv5 function - rumidas package - RDocumentation, (CRAN).

Fannie Mae, 2025, Refinance Application-Level Index (RALI), (Fannie Mae).

Federal Reserve Bank of St. Louis, 2022, TED Spread (DISCONTINUED), (FRED).

Financial Industry Regulatory Authority (FINRA), 2025, Margin Statistics, (FINRA).

Ice Data Indices, LLC, 2025, ICE BofA US High Yield Index Option-Adjusted Spread, (FRED).

Moshkelgosha, 2024, S&P 500 Bullish Percent Index: Cyclical Behavior, (TradingView).

National Association of Active Investment Managers, 2025, NAAIM Exposure Index, (NAAIM).

S&P Dow Jones Indices, 2025, S&P 500 Index Data, (S&P Global).

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson, 2020, Daily News Sentiment Index, (Federal Reserve Bank of San Francisco).

TradingView, 2025, USI:PCSPX – Put/Call Ratio SPX Chart, (TradingView).

University of Michigan, 2025, Surveys of Consumers Charts, (University of Michigan).

U.S. Census Bureau, 2025, Business Formation Statistics – Weekly Data, (U.S. Census Bureau).

Woolley, John T., and Gerhard Peters, 2024, Presidential Job Approval—All Data, (The American Presidency Project).

Appendix

Tables

Table A.1: Pairwise Correlations > $|0.5|$

Variable 1	Variable 2	Correlation
INST_MONETARY_UNCERTAINTY	INST_POLICY_UNCERTAINTY	0.59
INV_MARGINDEBT	INV_BPI	0.53
INV_EXPOSURE	INV_BPI	0.52

Table A.2: Variance Inflation Factors (VIF)

Variable	VIF Value	Potential Multicollinearity (> 5)
logRV_lag_biweek	33.48	Yes
logRV_lag_week	32.60	Yes
logRV_lag_month	16.20	Yes
logVIX	13.59	Yes
logRV_lag	7.41	Yes
logRV	6.96	Yes
INV_EXPOSURE	4.15	No
INST_POLICY_UNCERTAINTY	3.03	No
INST_TED	2.72	No
INV_BEAR	2.49	No
INST_MONETARY_UNCERTAINTY	2.42	No
INV_BPI	2.29	No
INST_PRES_APP	1.99	No
INV_SKEW	1.93	No
INV_BULL	1.93	No
jump	1.86	No
INV_MARGINDEBT	1.70	No
INST_HYBS	1.49	No
daily_return	1.44	No
INV_EXPOSURE_STD	1.39	No
CON_NEW_BUSINESS	1.35	No
INST_GEOPOL_RISK	1.31	No
CON_REMORTGAGE	1.29	No
CON_SENT	1.27	No
INV_PCSPX	1.07	No
INST_NEWS_SENT	1.06	No

Table A.3: Wald Test Results

Variable	F Statistic	p-value
INV_SKEW	61.982	0.000
INV_PCSPX	0.849	0.357
INV_BPI	28.213	0.000
INV_BULL	43.438	0.000
INV_BEAR	21.416	0.000
INV_EXPOSURE	16.143	0.000
INV_EXPOSURE_STD	5.925	0.015
INV_MARGINDEBT	4.451	0.035
CON_SENT	22.437	0.000
CON_NEW_BUSINESS	10.578	0.001
CON_REMORTGAGE	0.275	0.600
INST_HYBS	3.271	0.071
INST_TED	8.246	0.004
INST_POLICY_UNCERTAINTY	54.062	0.000
INST_GEOPOL_RISK	140.731	0.000
INST_MONETARY_UNCERTAINTY	4.144	0.042
INST_NEWS_SENT	0.578	0.447
INST_PRES_APP	178.276	0.000

Note: Covariance: Newey–West (HC1 adjustment).

Table A.4: PCA Loadings for Sentiment Variables

	PC1	PC2	PC3	PC4	PC5
INV_SKEW	0.331	0.127	0.048	-0.162	0.074
INV_PCSPX	0.021	0.110	-0.172	0.072	-0.672
INV_BPI	0.269	-0.384	0.030	-0.121	-0.134
INV_BULL	0.066	-0.382	-0.342	-0.345	-0.127
INV_BEAR	0.298	0.235	0.327	0.164	0.020
INV_EXPOSURE	0.387	-0.085	-0.098	-0.216	-0.019
INV_EXPOSURE_STD	-0.170	0.074	0.107	0.510	-0.220
INV_MARGINDEBT	0.219	-0.357	-0.033	0.152	-0.034
CON_SENT	0.335	0.337	-0.038	-0.065	0.028
CON_NEW_BUSINESS	0.191	0.323	0.070	-0.253	0.032
CON_REMORTGAGE	-0.080	0.215	0.177	-0.398	-0.032
INST_HYBS	-0.073	0.189	-0.154	-0.184	-0.391
INST_TED	-0.307	0.155	-0.354	-0.009	0.229
INST_POLICY_UNCERTAINTY	-0.284	-0.143	0.388	-0.330	-0.060
INST_GEOPOL_RISK	0.227	0.244	-0.040	0.086	0.035
INST_MONETARY_UNCERTAINTY	-0.324	0.112	0.168	-0.312	0.011
INST_NEWS_SENT	0.036	-0.136	-0.015	0.050	0.471
INST_PRES_APP	0.061	-0.219	0.595	0.034	-0.151

Variables were scaled prior to PCA (mean=0, sd=1).

Table A.5: Robustness Results: VIX-HAR-X

	Dependent variable:
	logVIX Lead
logVIX	0.897***
	(0.036)
logVIX_lag	0.003
	(0.040)
logVIX_lag_week	−0.041
	(0.053)
logVIX_lag_biweek	0.037
	(0.041)
logVIX_lag_month	0.026
	(0.023)
jump	0.020**
	(0.009)
daily_return	0.408**
	(0.200)
INV_SKEW	−0.001
	(0.002)
INV_PCSPX	0.004**
	(0.002)

Continued on next page

Table A.5 – continued from previous page

	Dependent variable:
	logVIX Lead
INV_BPI	−0.013***
	(0.002)
INV_BULL	0.004*
	(0.002)
INV_BEAR	−0.006***
	(0.002)
INV_EXPOSURE	−0.002
	(0.003)
INV_EXPOSURE_STD	−0.0003
	(0.001)
INV_MARGINDEBT	−0.007***
	(0.002)
CON_SENT	−0.001
	(0.001)
CON_NEW_BUSINESS	−0.001
	(0.001)
CON_REMORTGAGE	0.003*
	(0.001)
INST_HYBS	−0.001
	(0.002)
INST_TED	0.003
	(0.002)
INST_POLICY_UNCERTAINTY	0.010***
	(0.002)
INST_GEOPOL_RISK	−0.004
	(0.005)
INST_MONETARY_UNCERTAINTY	−0.003
	(0.002)
INST_NEWS_SENT	0.001
	(0.001)
INST_PRES_APP	0.009***
	(0.003)
Constant	0.222***
	(0.029)
Observations	3,489
R^2	0.964
Adjusted R^2	0.963
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors computed via Newey–West

Table A.6: HAR: Full vs Subsample Estimates

	Dependent variable:											
	Log Realised Variance											
	Base-HAR (Full)	Base-HAR (1st Half)	Base-HAR (2nd Half)	HAR (Full)	HAR (1st Half)	HAR (2nd Half)	HAR-X (Full)	HAR-X (1st Half)	HAR-X (2nd Half)	HAR-X-LASSO (Full)	HAR-X-LASSO (1st Half)	HAR-X-LASSO (2nd Half)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
logRV	0.449***	0.354***	0.532***	0.315***	0.240***	0.331***	0.277***	0.206***	0.276***	0.277***	0.206***	0.294***
	(0.026)	(0.035)	(0.036)	(0.027)	(0.037)	(0.035)	(0.025)	(0.037)	(0.032)	(0.026)	(0.037)	(0.034)
logRV_lag	0.137***	0.144***	0.115***	0.113***	0.117***	0.081**	0.098***	0.101***	0.055*	0.098***	0.098***	0.066**
	(0.024)	(0.033)	(0.035)	(0.023)	(0.031)	(0.034)	(0.022)	(0.029)	(0.031)	(0.023)	(0.030)	(0.033)
logRV_lag_week	0.227***	0.241***	0.199***	0.192***	0.221***	0.137**	0.170***	0.193***	0.099	0.172***	0.199***	0.114*
	(0.054)	(0.078)	(0.075)	(0.048)	(0.066)	(0.066)	(0.050)	(0.066)	(0.066)	(0.049)	(0.066)	(0.068)
logRV_lag_biweek	−0.012	0.064	−0.032	−0.042	0.056	−0.137*	−0.080	−0.046	−0.109	−0.071	0.004	−0.136*
	(0.060)	(0.089)	(0.080)	(0.056)	(0.075)	(0.070)	(0.056)	(0.072)	(0.067)	(0.055)	(0.073)	(0.071)
logRV_lag_month	0.151***	0.142***	0.110**	−0.030	−0.107*	0.017	−0.062	−0.205***	−0.050	−0.064	−0.223***	0.002
	(0.041)	(0.054)	(0.055)	(0.038)	(0.058)	(0.048)	(0.044)	(0.074)	(0.051)	(0.043)	(0.069)	(0.050)
jump				−0.033	0.100	−0.235**	−0.045	0.037	−0.066	−0.036	0.055	−0.073
				(0.060)	(0.073)	(0.111)	(0.064)	(0.075)	(0.104)	(0.063)	(0.076)	(0.098)
daily_return				−6.902***	−6.881***	−5.405**	−4.098***	−4.222***	−1.936	−4.267***	−4.521***	−2.534
				(1.260)	(1.094)	(2.458)	(1.165)	(1.030)	(2.361)	(1.166)	(1.047)	(2.373)
logVIX				1.196***	1.035***	1.848***	1.386***	1.328***	2.119***	1.353***	1.216***	1.960***
				(0.105)	(0.135)	(0.155)	(0.117)	(0.173)	(0.161)	(0.120)	(0.173)	(0.158)
INV_SKEW							−0.077***	−0.066**	−0.057***	−0.069***	−0.057**	−0.042**
							(0.015)	(0.030)	(0.017)	(0.013)	(0.027)	(0.016)
INV_PCSPX							0.007	0.002	0.014			
							(0.010)	(0.011)	(0.022)			
INV_BPI							−0.080***	−0.098***	−0.092***	−0.083***	−0.107***	−0.087***
							(0.018)	(0.026)	(0.025)	(0.017)	(0.024)	(0.023)
INV_BULL							−0.008	−0.052**	0.056**			
							(0.019)	(0.022)	(0.028)			
INV_BEAR							−0.037**	−0.064**	0.042	−0.027*	−0.039	0.028
							(0.017)	(0.025)	(0.029)	(0.016)	(0.025)	(0.024)
INV_EXPOSURE							−0.029	−0.040	−0.029	−0.017	−0.049**	−0.005
							(0.019)	(0.026)	(0.033)	(0.017)	(0.023)	(0.028)
INV_EXPOSURE_STD							−0.015	−0.027	−0.006			
							(0.011)	(0.018)	(0.017)			
INV_MARGINDEBT							−0.020	−0.033*	−0.023	−0.020	−0.029	−0.018
							(0.014)	(0.019)	(0.017)	(0.014)	(0.019)	(0.017)
CON_SENT							0.009	0.019	−0.049**			
							(0.014)	(0.014)	(0.022)			
CON_NEW_BUSINESS							0.026**	0.026	0.043***			
							(0.013)	(0.021)	(0.016)			
CON_REMORTGAGE							−0.001	0.006	−0.004			
							(0.015)	(0.016)	(0.021)			
INST_HYBS							0.040***	0.034**	0.024	0.040***	0.033**	0.029
							(0.011)	(0.014)	(0.020)	(0.011)	(0.013)	(0.020)
INST_TED							−0.035*	−0.004	−0.131***	−0.034*	0.005	−0.116**
							(0.019)	(0.017)	(0.049)	(0.019)	(0.017)	(0.050)
INST_POLICY_UNCERTAINTY							−0.030	0.004	0.019	−0.028*	0.008	−0.040
							(0.020)	(0.026)	(0.043)	(0.017)	(0.023)	(0.027)
INST_GEOPOL_RISK							0.008	−0.038	0.061			
							(0.039)	(0.083)	(0.043)			
INST_MONETARY_UNCERTAINTY							−0.003	0.002	−0.087**			
							(0.022)	(0.028)	(0.035)			
INST_NEWS_SENT							0.007	0.019	−0.00002			
							(0.011)	(0.014)	(0.015)			
INST_PRES_APP							−0.022	−0.012	0.048	−0.024	0.007	−0.036
							(0.023)	(0.033)	(0.043)	(0.024)	(0.030)	(0.040)
Constant	−0.472***	−0.526***	−0.802***	−7.971***	−7.655***	−11.002***	−9.932***	−11.250***	−13.384***	−9.732***	−10.568***	−12.264***
	(0.098)	(0.151)	(0.208)	(0.650)	(0.965)	(0.879)	(0.678)	(1.269)	(0.915)	(0.681)	(1.286)	(0.879)
MSE	0.393	0.382	0.397	0.354	0.346	0.347	0.341	0.328	0.332	0.342	0.331	0.337
Observations	3,511	1,755	1,756	3,511	1,755	1,756	3,511	1,755	1,756	3,511	1,755	1,756
R^2	0.750	0.709	0.685	0.775	0.737	0.724	0.783	0.750	0.737	0.783	0.748	0.733
Adjusted R^2	0.750	0.709	0.684	0.775	0.736	0.723	0.782	0.746	0.733	0.782	0.746	0.730

Note: *p<0.1; **p<0.05; ***p<0.01

Plots

Figure A.1: Daily log-Realised-Variance of the S&P 500

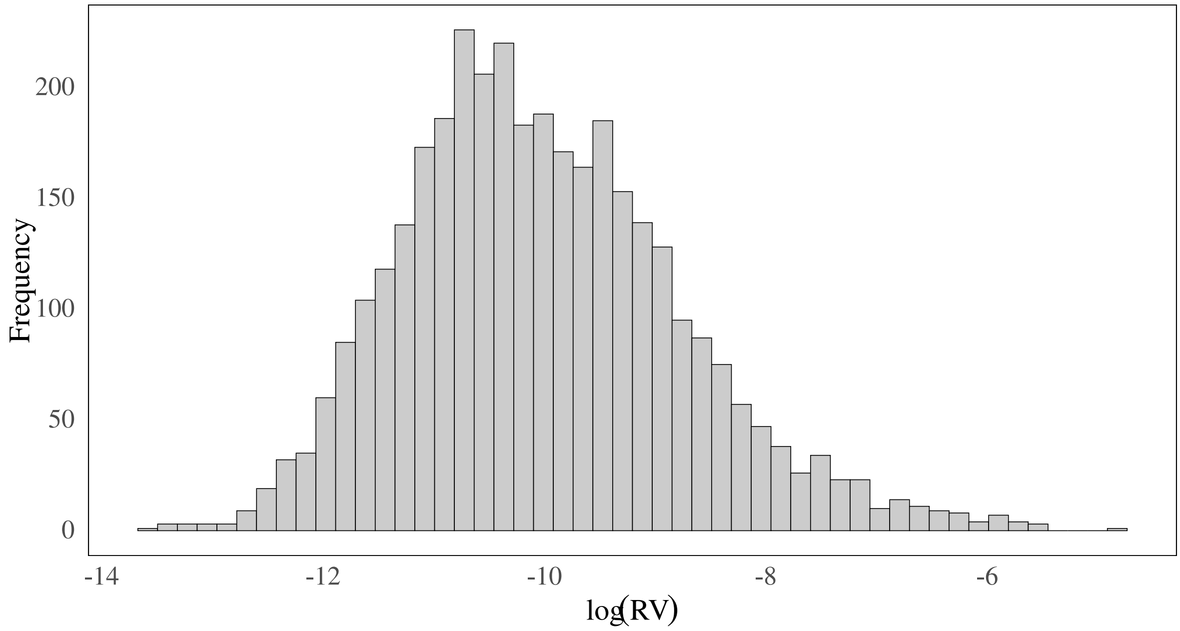


Figure A.2: Daily log-Realised-Variance of the S&P 500

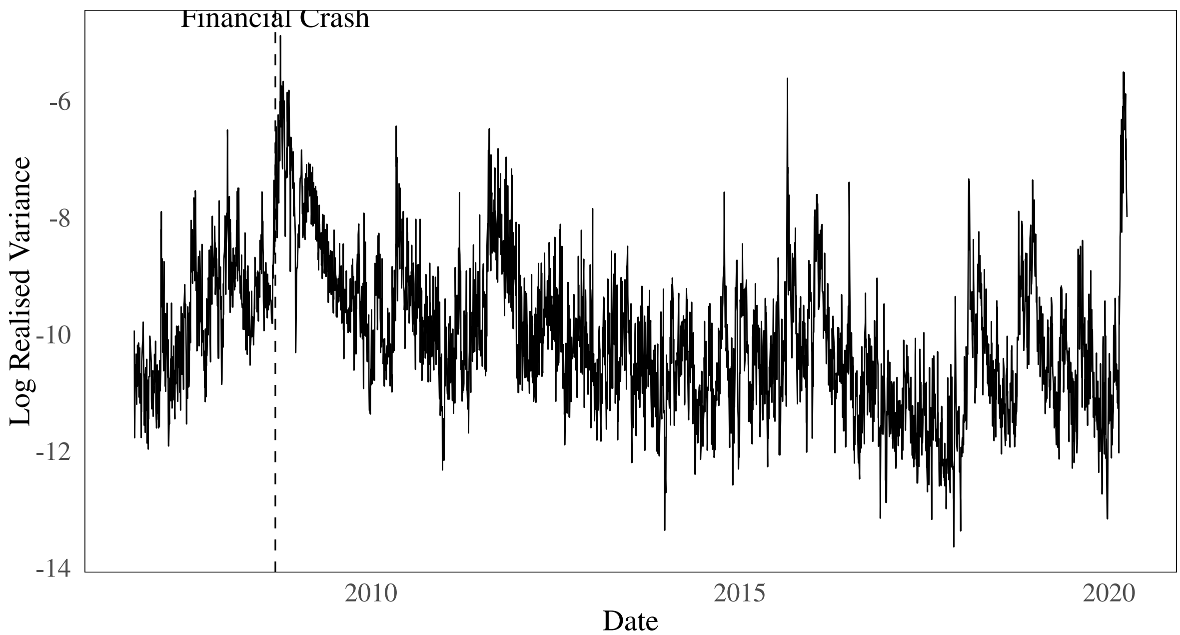


Figure A.3: Sentiment Variable Boxplots

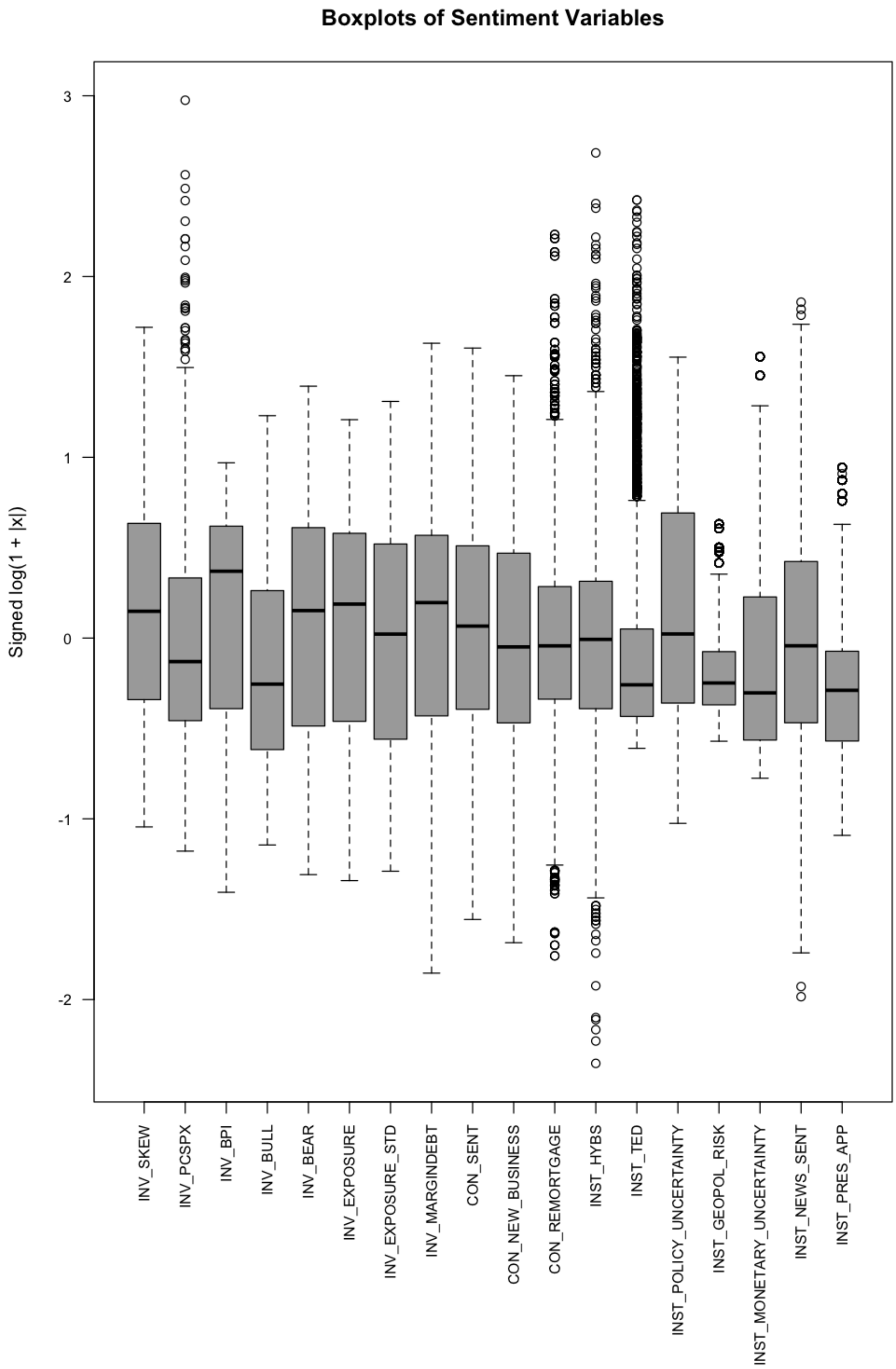


Figure A.4: Correlation Matrix

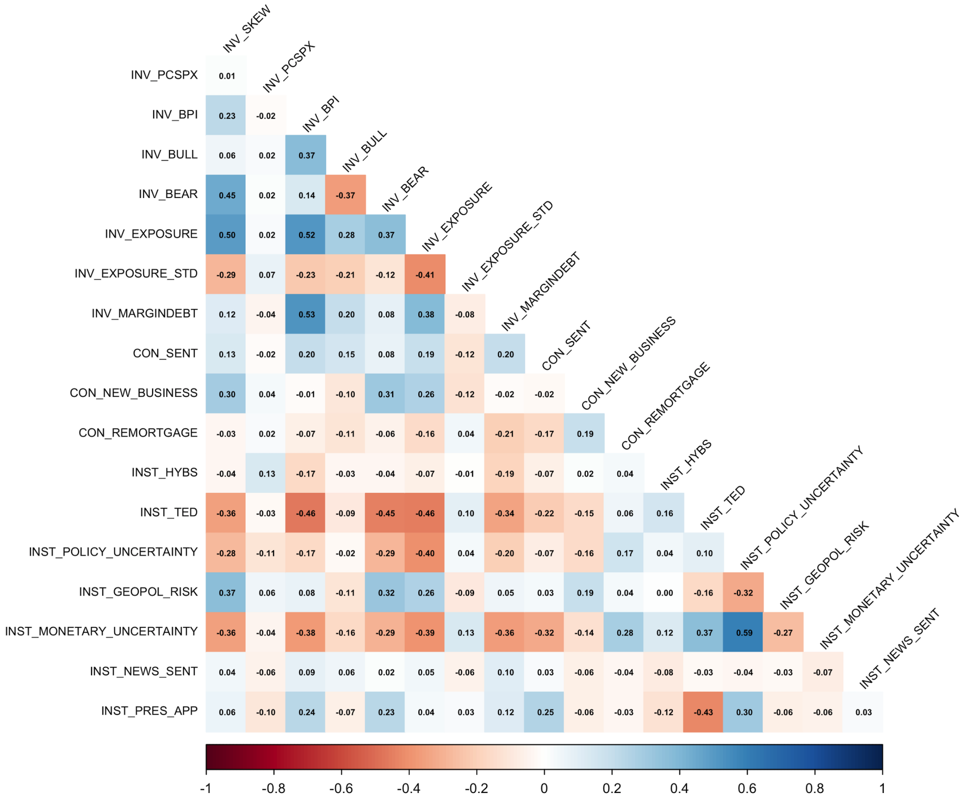


Figure A.5: Cumulative Variance Explained by Principal Components

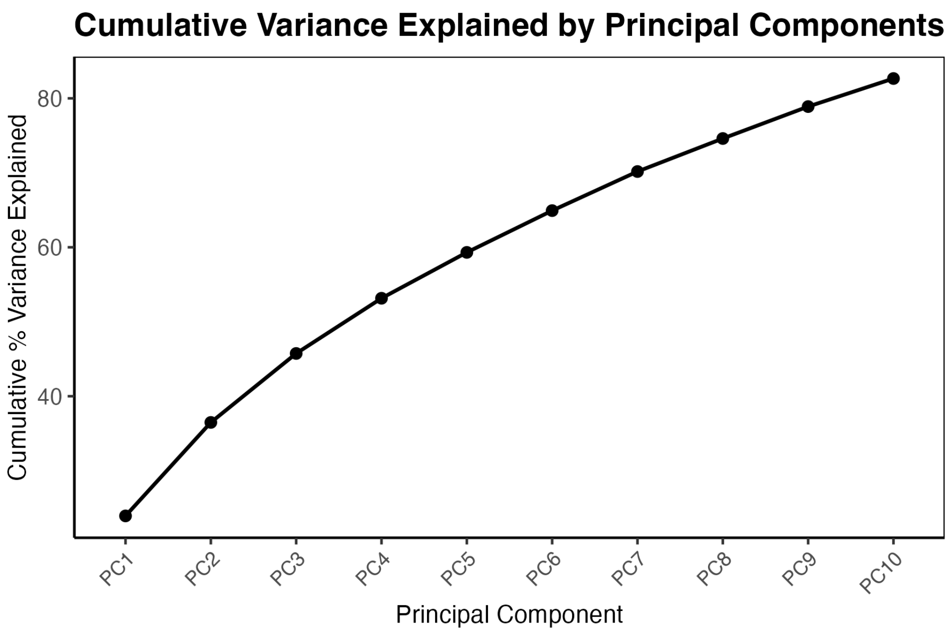


Figure A.6: ACF Plot

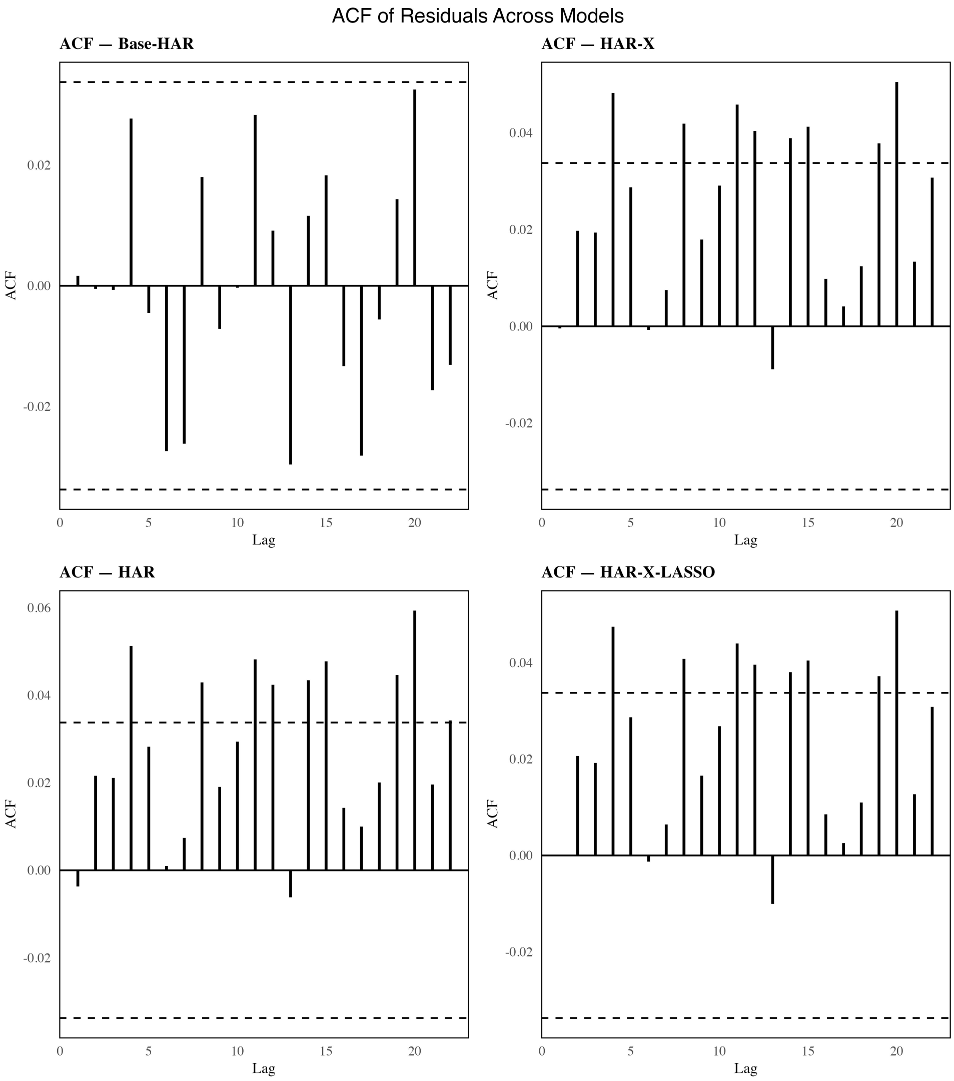


Figure A.7: Q–Q Plot of log(RV)

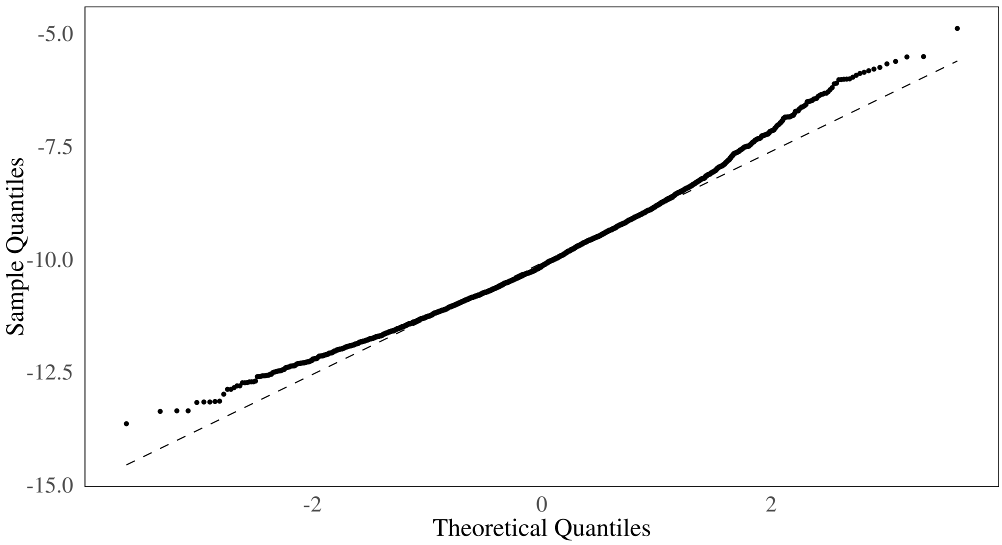


Figure A.8: Cumulative Forecast Error

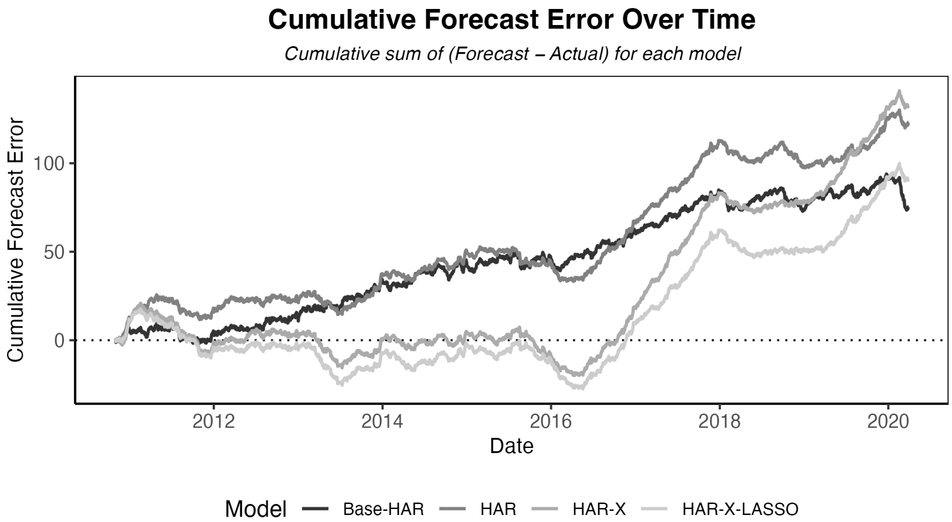
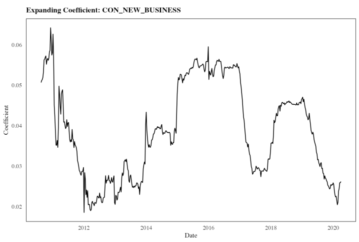


Figure A.9: Expanding Coefficient Stability Plots for Sentiment Variables

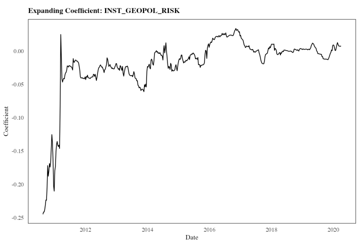
(a) CON_NEW_BUSINESS



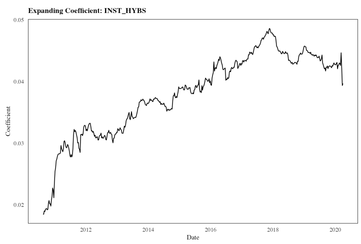
(b) CON_REMORTGAGE

(c) CON_SENT

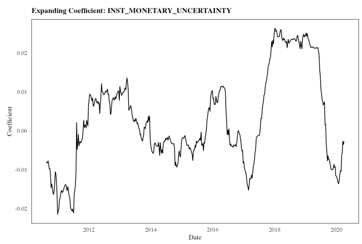
(d) INST_GEOPOL_RISK

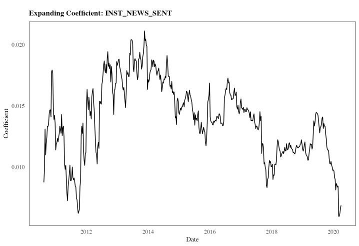


(e) INST_HYBS

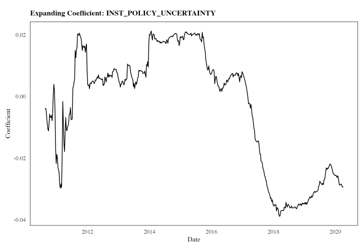


(f) INST_MONETARY_UNCERTAINTY

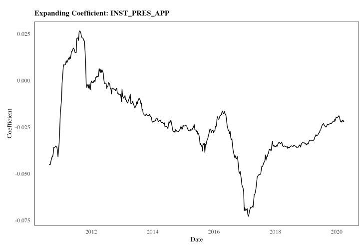




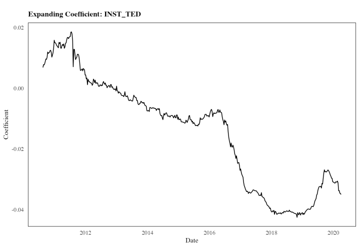
(g) INST_NEWS_SENT



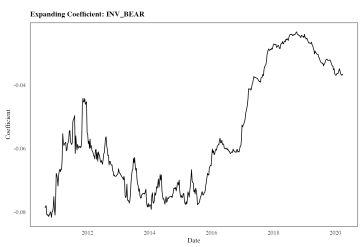
(h) INST_POLICY_UNCERTAINTY



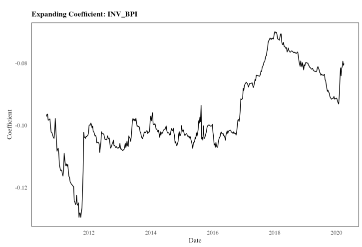
(i) INST_PRES_APP



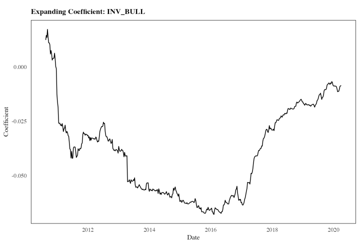
(j) INST_TED



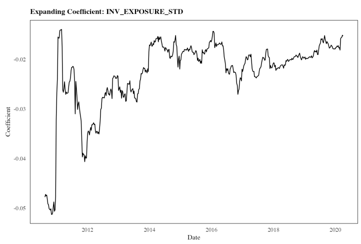
(k) INV_BEAR



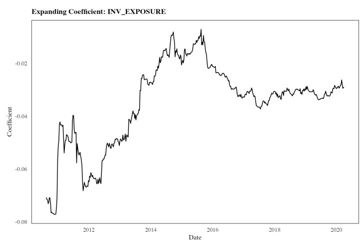
(l) INV_BPI



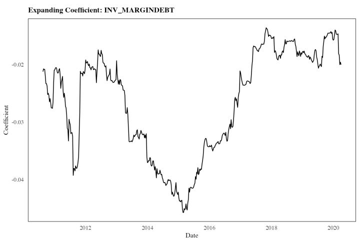
(m) INV_BULL



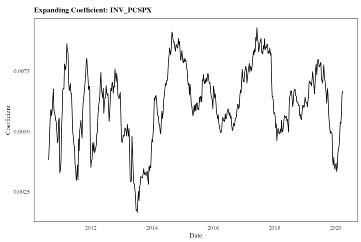
(n) INV_EXPOSURE_STD



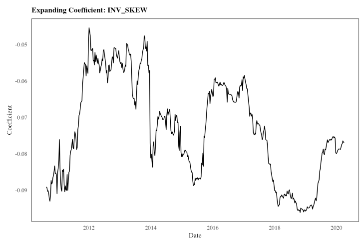
(o) INV_EXPOSURE



(p) INV_MARGINDEBT



(q) INV_PCSPX



(r) INV_SKEW

Other

AI Appendix

The only tool used for this thesis within Generative AI was ChatGPT. The use of ChatGPT served two purposes. The first purpose was to serve as an advisor and guide for the creation and development of the R code. No code was actively copied, as we also saw that its outputs were inconsistent when suggesting improvements. However, the usage of it for proofreading purposes enabled us to test whether the code we produced gave us the output that we intended and to optimise its performance. The second purpose of ChatGPT was to proofread parts of the text for clarity and when we needed inspiration on how to restructure complicated sentences. As ChatGPT more often than not tends to neglect information when reformatting, we refrained from actively including the output, but used its suggestions and ideas for improvement for inspiration. Since the paper was written using LaTeX, ChatGPT was also used for assistance regarding formatting questions in the LaTeX program.