Author:
Yang, Chen

Title:
Predicting volatility with Twitter sentiment
an application to the US stock market.

Predicting volatility with Twitter sentiment: an application to the US stock market.

Chen Yang

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of MPhil Accounting and Finance in the Faculty of Business school.

Word count: 26628

Table of Content

Abstract

This dissertation investigates whether individual tweets related to the S&P500 Index can predict volatility in future returns. A sample of 3,329,267 tweets containing the keyword “SPX” was collected from the period 2012 to 2021. We applied Principal Component Analysis (PCA) to reduce the dimensionality of the word frequency data and then integrated it with the Heterogeneous Autoregressive (HAR) model. We evaluated the in-sample and out-of-sample forecasting performance of various HAR-PCA models using different estimation window schemes and compared them with the original HAR model. We found that HAR-PCA models generally outperform the HAR model, especially during periods of particularly high and low volatility. Our findings demonstrate the economic relevance of HAR-PCA models for portfolio investment and contribute to the literature by linking investor sentiment to return volatility using a word-based method, which avoids the complications of applying advanced algorithms.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....  ..... DATE: 10/12/2023

Chapter 1. Introduction

Volatility modelling and forecasting are essential tasks in risk management, derivative pricing, portfolio selection, and central bank policymaking. Accurate volatility forecasts facilitate decision-making by market participants and market regulators. However, predicting volatility constitutes a demanding job. There are several variables that can be considered in predicting volatility but, in this dissertation, we concentrate on investor sentiment as expressed on the social media platform, Twitter. Historically, investors have relied on information intermediaries (e.g., professional financial news sources and financial adviser services) to obtain timely information and data about the prospects for financial markets. However, during the past decade, many new information resources have emerged that are readily available to anyone involved in the financial markets. With the proliferation of online financial information, especially on social media platforms, individual investors have increasingly turned to internet-based social media to capture and disseminate relevant information. The most revolutionary development in the distribution of information in this way has been the rise and growth of social networking websites such as Twitter, which enable market players to communicate their opinions on the financial markets in real time and which provide vast forums for discussion.

As an intriguing and fast-growing new resource for information on financial markets, the primary benefit of Twitter is that it mitigates information asymmetry (Blankespoor, Miller and White, 2014). Institutional market participants and professional financial services companies used to have better and quicker access to financial information than individual investors. In contrast, because Twitter's user base is highly diverse, it offers non-professional participants seeking market information significant advantages over more homogeneous specialist social networking sites and conventional information gatekeepers such as financial analysts. Twitter provides a platform for investors to share and exchange messages and opinions about the market. As a result, individuals potentially have better prospects and the ability to make more informed decisions about future events. Extensive research, particularly in textual analysis, has been conducted to examine the flow of sentiment in the financial markets. However, most of these studies have concentrated on analysing news articles, annual reports, and 10-K filings (Antweiler and Frank (2004), Das and Chen (2007), Tetlock (2007), Loughran and McDonald (2011), Jegadeesh and Wu (2013), Purda and Skillicorn (2015), etc. There has been only limited coverage of social media text. The widely used dictionary, developed by Loughran and McDonald (2011) to classify

the tone of financial texts, is derived from 10-K filings. There is no evidence that it performs well in analysing short tweets written in informal language. This dissertation adds to the literature by investigating whether or not individual tweets about the stock market can be used to predict the volatility of stock returns. Specifically, we define the realized volatility of the S&P500 Index as the dependent variable and explore the following question: Do occurrences of certain words in individual tweets regarding the S&P500 Index predict the realized volatility of the S&P500 Index?

To explore this research question, we collected a sample of 3,329,267 tweets posted over a period of nine years (1 January 2012 to 31 December 2021). All of these tweets are directly related to the S&P500 Index because we only selected tweets that contained the keyword "SPX," which is the ticker of the S&P500 Index. Texts that use the more generic term "S&P500" contain a great deal of noise, and frequently include irrelevant and sometimes misleading information. Use of "SPX" as a substitute for "S&P500 Index" is a widely accepted convention to filter out messages that have a wider scope than purely investment-related aspects of the market (Jiang and Tian (2005), Buccioli and Kokholm (2021), and Guyon (2020). After pre-processing the data, we analysed both in-sample and out-of-sample forecasting performance by splitting the sample period into an in-sample period (1 January 2012 to 31 December 2013) and an out-of-sample period (1 January 2014 to 31 December 2021). Firstly, we used a bag-of-words approach to represent the corpus based on the appearance of individual words. Then, we used Principal Component Analysis (henceforth PCA), to summarize the textual content with a smaller set of variables to augment Corsi's (2009) Heterogeneous Autoregressive (HAR) model. PCA reduces dimensionality by converting a set of correlated variables into a smaller set of uncorrelated variables. These components capture most of the variability while reducing noise and redundancy. For instance, the first four components captured 68% of the total variability of the series in the original dataset. Our study evaluates eight HAR-PCA models, including the original HAR model's independent variables (daily, weekly, and monthly lagged realized volatility) and a maximum of eight of the principal components that explain most of the variability in the daily word frequency data.

We assessed the predictive accuracy of the various models both in-sample and out-of-sample (OOS). For the OOS analysis, we employed two estimation window schemes: rolling window and expanding window. The HAR model served as a benchmark to compare the forecasting performance of the sentiment-augmented predictive models. Using this methodology, we linked investor sentiment to

return volatility. This word-based method has an advantage over previous studies in avoiding the complications of applying advanced algorithms. We focus on the most basic unit—words—to capture fluctuations in investor sentiment. Also, unlike previous research that employed pre-specified and fixed word lists, our approach does not rely on a dictionary (Tetlock, Saar-Tsechansky and Macskassy (2008) and Loughran and McDonald (2011). Instead, our principal component variables change dynamically as the sample period rolls forward or expands.

Our findings can be summarized in three key observations. First, the HAR-PCA models generally outperform the HAR model in terms of their capability to forecast future volatility. The most effective model for the full sample period was found to be the daily-updated log-HAR-PCA-5 model (incorporating the top five principal components), using a rolling window approach. Interestingly, while log-HAR-PCA-5 is the winning model, adding principal components does not dramatically alter the performance of the model. In fact, adding more than five components reduces the OOS performance of the model. Second, the performance discrepancy between the HAR-PCA and HAR models is more pronounced during periods of high and low volatility when compared to periods of moderate volatility. This suggests that word frequency data have greater significance and applicability for predicting volatility during periods marked by particularly high or low market uncertainty. It was also found that the statistical properties of our models remain robust when forecasting an alternative volatility measure: realized kernel. Third, the HAR-PCA models demonstrate economic relevance when utilized for portfolio investment, achieving higher utility gains in comparison with the original HAR model. This finding highlights the potential benefits of applying HAR-PCA models to forecast volatility when the goal is to build an investment strategy based on this input, as the HAR-PCA models can lead to better-informed decision-making and improved financial outcomes.

Our method allows for replicability, refinement by future researchers, and a more transparent analysis process, avoiding the "black box" issue that is a feature of complex machine learning algorithms. The approach also offers a baseline model for experts in various fields to apply their knowledge, potentially enhancing results and fostering a more comprehensive understanding of volatility. The dissertation is structured as follows. Section 2 is devoted to a discussion of the existing literature. In Section 3, we describe the data. Section 4 details the research methodology, while Section 5 provides the empirical and supplementary results. Section 6 describes the robustness checks carried out on our statistical results,

and Section 7 explores the economic significance of our models. Section 8 presents our conclusions.

Chapter 2. Literature review

2.1 Investor's sentiment and stock market

Investors' sentiment plays a significant role in forecasting stock return volatility and liquidity. The media influence investors' decisions. Conversely, investors also express their opinions through social media to affect others. Many studies have summarized the prediction power of media coverage in the stock market. At the same time, there is yet to be a consensus about how explicitly the stock return volatility relates to investors' sentiment. As Keynes noted in 1937, because of investors' "animal spirits," market prices often fluctuate wildly and for no apparent reason. De Long et al. (1990) demonstrate that increased noise trading, more mispricing, and more volatility are all shown to result from changes in market sentiment if unaware noise traders execute their trading strategies relying on sentiment. Based on these works, many researchers seek to establish the connection between sentiment and stock market returns (Tumarkin and Whitelaw (2001), Baker and Wurgler (2006), Zhang, Fuehres and Gloor (2011), Chung, Hung and Yeh (2012), Boudoukh et al. (2013), Oliveira et al. (2013), Liew and Budavári (2016), Sun, Najand and Shen (2016)). For example, Chung, Hung and Yeh (2012) measure the sentiment using Baker and Wurgler's (2006) orthogonalized sentiment index. They document that sentiment has predictive power for return in the expansion state, but not in recession. Boudoukh et al. (2013) argue that the news's impact on the stock price changes more noticeably when the news's type and tone are appropriately detected. In a more recent study, Sun, Najand, and Shen (2016) examined the predictive power of the Thomson Reuters MarketPsych Index, a sentiment indicator that takes into account newswires, online news, and social media to forecast 30-minute returns. They found that changes in investor sentiment can forecast intraday stock returns.

The following works are similar: the authors use various measures of investor sentiment, such as Google search data (Gao, Ren, and Zhang, 2019), a new sentiment index (Huang et al., 2015), social media messages (Kurov, 2010), and proxies such as implied volatility index (VIX) (McLean and Zhao, 2014). Studies on investor sentiment reveal a link between investor sentiment and stock returns, with higher levels of sentiment leading to greater profits from exploiting market anomalies (Zouaoui, Nouyrigat and Beer, 2011), increased investments (Arif and Lee, 2014) and increased sensitivity to earnings news (Mian and Sankaraguruswamy, 2012). Furthermore, Renault (2017) investigated how investment

decisions relating to proxies for stock market mispricing; findings suggested that firms with high R&D intensity or turnover were more likely affected by mispricing (Sun, Najand and Shen, 2016). This indicates that these companies are more sensitive to shifts in investor sentiment and could be more vulnerable to market volatility as a result of such changes. Previous research additionally examined how investors' sentiments might predict stock market crashes (Polk and Sapienza, 2004). Their results suggest that sentiment raises the probability of crises in countries with a tendency towards herd behavior and overreaction or countries with low institutional involvement (Smales, 2017). This highlights the importance of considering the broader societal context when evaluating the potential impact of investor sentiment on stock markets and underscores the need for careful analysis of market conditions in order to mitigate the risk of market instability. Another key finding of the studies is that investor sentiment plays a central role in the stock market (Firth, Wang and Wong, 2015). The research demonstrates that sentiment drives noise trading, which can significantly impact stock prices and market returns (Zhou, 2018). The authors argue that sentiment is a key driver of market behavior and can profoundly impact investment decisions, particularly in the short term.

Besides, in Stambaugh and Yuan (2017), a four-factor asset pricing model is proposed, which accommodates a wide range of anomalies than previous models and takes into account the influence of investor sentiment. In another study (Daniel, Hirshleifer and Sun, 2019), a factor model is presented that aims to explain the cross-section of US equity returns by incorporating investor psychology and the long and short-horizon mispricing caused by managers' decisions and investor inattention respectively. A deep learning-based stock market prediction model (Jin, Yang and Liu, 2020) is also proposed, which considers investor sentiment to improve the prediction accuracy and overcome the challenges of noise, volatility, and the complexity of stock price sequences by involving sentiment, decomposing the price sequence with Empirical Modal Decomposition (EMD), and using a revised Long short-term memory (LSTM) network with an attention mechanism. This has important implications for individual investors, highlighting the importance of being mindful of sentiment when making investment decisions. It also underscores the need for investors to be vigilant in monitoring market conditions and to be prepared to adjust their investments in response to shifts in investor sentiment.

Similar findings investigating the sentiment have been made by the following academics: Antweiler and Frank (2004); Ho, Shi, and Zhang (2013); Lee, Hutton and Shu (2015); Da, Engelberg and Gao (2015); See-To and Yang (2017); Siganos, Vagenas-Nanos and Verwijmeren (2017); Behrendt and Schmidt (2018); Rahimikia and Poon (2020). For instance, Antweiler and Frank (2004) found that stock market volatility may be predicted by analyzing the tone of comments made on Yahoo! Finance and Raging Bull. According to research by Ho, Shi and Zhang (2013), the intraday volatility of certain US equities is highly impacted by news sentiments. Lee, Hutton and Shu (2015) illustrate that sentiment can be priced as a systematic risk, and the changes in sentiment can lead to revisions in volatility and excess return. Da, Engelberg and Gao (2015) aggregated consumer Internet search activity to create the "FEARS" index, which measures investor sentiment. This measure can forecast sudden increases in volatility and short-term return reversals. In a more recent study, Rahimikia and Poon (2020) found that including a limit order book and news sentiment in the HAR model can improve volatility forecasting. Most studies support the idea that news or social media sentiment can or at least assist in predicting stock market return and volatility (e.g., Baker and Wurgler (2006), Chung, Hung and Yeh (2012), Boudoukh et al. (2013), while others are skeptical (e.g., Antweiler and Frank (2004), Oliveira et al. (2013)). Generally, consensus has yet to be reached on forecasting stock market returns or volatility using sentiment analysis. The most recent topic is not whether investor sentiment affects stock market pricing but how to assess and quantify the effects of investor sentiment.

Overall, the studies shed light on the impact of investor sentiment on stock markets and stock returns. The authors argue that sentiment is a key driver of market behavior and can significantly impact investment decisions and market outcomes. However, most of the papers employing semantic analysis utilize Loughran-McDonald word lists that are unsuitable for Twitter. Our research addresses this gap in the literature by proposing a method for directly assessing sentiment from text.

2.2 Measuring the investor sentiment using textual analysis

2.2.1 Techniques

To build a link between investor sentiment and capital markets, the first problem that needs to be tackled is sentiment. Textual analysis is an area of research utilizing computational and statistical methods to extract meaningful information from text data.

Manual classification

Prior to the era of big data, researchers analyse the text of financial disclosure using the manual classification method. For example, Botosan (1997) manually identifies the disclosure quality using a score based on 35 financial and nonfinancial information items. Similarly, Bryan (1997) applied a scoring system to manually classify the MD&A (Management Discussion and Analysis) disclosures into numerical variables. He assigned a score of +1, 0, or -1 to each disclosure based on whether it indicated a positive, neutral, or negative effect on future performance. More recently, Li (2010) examined the tone of forward-looking statements (FLS) using a combination of manual classification, and Naive Bayesian supervised machine learning algorithm. He finds that the tone of FLS has explanatory power to key company variables, such as the current performance, accruals, firm size, and return volatility. However, manual analysis is also subject to some limitations. Firstly, it can be a time-consuming and labor-intensive process, especially when dealing with a large volume of data. This can lead to inconsistencies in analysis, as different analysts may interpret the same information differently. Secondly, manual analysis is also susceptible to human biases and errors, such as misinterpretation of the text or data entry mistakes. This can affect the accuracy and reliability of the analysis, potentially leading to incorrect or incomplete conclusions. Increasing computer processing speed makes textual analysis more feasible, but the overwhelming amount of unstructured data made available by internet archives and social media sites is also crucial. With the development of natural language processing, text information can now be transformed into quantitative information more efficiently. Two general types of text analysis approaches dominate the accounting and finance literature. The first is the dictionary or bag-of-words approach, and the other is the machine learning approach.

Bag-of-words

Researchers employ bag-of-words methods to classify words based on predefined rules. In past studies, two lexicons were utilized almost exclusively to create sentiment assessments. The first is the Harvard Psychosociological Dictionary, which categorizes terms according to whether they are positive or negative to determine the tone of a document or news. Tetlock (2007) applies the General Inquirer text analysis tool to quantify the tone of the "Abreast of the Market" Wall Street Journal column according to the Harvard-IV dictionary. His result suggests that the media contains essential information about the stock market. Tetlock, Saar-Tsechansky and Macskassy (2008) then studied whether the tone of

words can predict stock return. The results support their hypothesis: negative words can predict low firm stock returns on the next trading day. Both of the above studies evaluate the tone of financial text using the Harvard-IV dictionary that developed within the domains of psychology and sociology. However, significant limitations exist on how the negative-sentiment Harvard-IV word list may be used in financial disclosures. An individual word in the English language may have multiple meanings depending on the context in which it is used. The Harvard Dictionary has a number of flaws, one of which is that it was not initially constructed with financial issues in mind. According to research by Loughran and McDonald (2011), the Harvard-IV-4 dictionary misclassifies about 75% of all negative words. In order to overcome the inapplicability of the Harvard dictionary, Loughran and McDonald (2011) developed a word list that is specifically applicable to the language of financial disclosures. They analyze the link between the words and 10-K filing returns, trading volume, and return volatility. According to their research, sentiment measures using the vocabulary classified by Loughran and McDonald (2011) are superior for capturing returns related to the 10-K. As a result, a significant number of following research in accounting and finance have adopted sentiment metrics based on the dictionary created by Loughran and McDonald (2011). (e.g., Law and Mills 2015). They made a key improvement by constructing a new dictionary that was more relevant to the economic context.

Many other word lists have been developed in addition to these two standard dictionaries. In order to catch unethical behavior by managers during earnings conference calls, Larcker and Zakolyukina (2012) compiled lists of both negative and positive words to use as red flags. Bodnaruk, Loughran and McDonald (2015) came up with a set of 184 terms to measure the depth of public companies' financial constraints. The list is compiled by investigating the vocabulary of at least 5% of all annual reports and selecting "tokens" that represent terms readers typically find constraining. They conclude their metric is more accurate than other constraint indices at predicting financial consequences like dividend omissions. Hope and Wang (2018) replicate Larcker and Zakolyukina's (2012) methodology and find that CEOs' big baths write-offs result in a significant widening of the company's bid-ask spreads. This demonstrates the viability of monitoring earnings conference calls for signs of managerial deceit. The economic policy uncertainty (EPU) measure developed by Baker, Bloom, and Davis (2016) is likewise based on expert judgment. They count how often certain words appear in major newspapers. As a result, their EPU index is associated with lower investment levels and higher stock return volatility. Moreover,

Soo (2018) developed a housing market sentiment index for 34 U.S. cities from 2000 to 2013 based on news articles. Her housing market sentiment indicator is proven to forecast future home price rises. Furthermore, this study concludes that an increase in the number of newspaper housing articles, including terms such as "highs" and "frenzied," is related to an increase in future house prices. More recently, Loughran, McDonald and Pragidis (2019) created a collection of 130 words expected to impact oil prices in order to analyze traders' propensity reflected through the news. They discover that the public tends to overreact in the near run to stories about oil. Words like "recovery," "trouble," and "assault" are among their most often used terms.

All the above literature applies the sentiment lexicon method, and these words lists are classified by expertise using their professional knowledge and experiences. While the subjectivity problem that the bag-of-words approach may face discourages researchers from defining a word list. Research has shown that new dictionaries, such as the García, Hu and Rohrer (2022), which incorporate bigrams models, can provide a more accurate representation of sentiment than traditional approaches.

Machine learning

Another strand of literature focuses on machine learning algorithms. Textual analysis in accounting has been gaining increasing attention as a research method due to its ability to process and analyze large amounts of data through natural language processing (NLP). NLP is a central component of textual analysis, and machine learning, especially deep learning, is emphasized as an important tool for NLP implementation. Accounting researchers are encouraged to increase their knowledge and use of machine learning in NLP (Bochkay et al., 2022). The use of NLP in financial analysis extends to the examination of social media activity and its impact on ETF premiums (Liu, 2023). Machine learning has been applied to extract and quantify firms' exposure to risk factors, which can be categorized into systematic and idiosyncratic risks and used to construct pricing factors (Lopez-Lira, 2023). The results suggest that production-based risks can be valuable in cross-sectional pricing and perform similarly to classic models in pricing a wide range of assets. However, it is important to note that commonly used platforms, such as Diction, may not be suitable for evaluating the tone of financial disclosures. A better tool for capturing the tone in the business text would be the Loughran-McDonald (2011) dictionary (Loughran and McDonald, 2015). Besides, the relationship between language used in corporate disclosures and increased AI readership has also been studied. Results indicate that firms adjust their

language to suit machine processing and avoid negative sentiment to be more favorable to algorithms (Loughran and McDonald, 2015).

Do humans or machines have an advantage when it comes to picking out the general tone of an extensive collection of financial statements by looking for key phrases? Some argue that employing computational approaches to generate word lists based on sentiment eliminates the subjectivity of human selection. The majority of newly developed approaches are derived from the field of machine learning and usually come under the categories of supervised or unsupervised learning. Supervised learning is utilized when a set of determination rules is accessible. In unsupervised approaches, researchers allow the algorithms, such as topic analysis, to explore hidden patterns in the data. Several existing studies investigate the application of machine learning in textual analysis. In its early application stages, machine learning in the accounting and finance literature relies on human assistance in text classification. (Huang et al. 2015). As mentioned above, Li (2010) manually classifies the degree to which forward-looking statements in the MD&A section are positive. Next, he feeds these manually categorized statements into a Naive Bayes Classifier (NBC). To maximize the possibility that the aggregate values would provide categorizations that match the manually specified groups, the NBC assigns scores to word patterns. Once the scores of the word patterns are determined, the classifier may assign confidence levels to assertions that have not yet been labeled. Combining human labor with machine learning has a key benefit over traditional text categorization methods in that the researcher may fine-tune the content analysis for a specific context. While there is an advantage to manual machine learning over purely manual categorization, it still has some of the same drawbacks regarding manual categorization.

In the more recent stage, newly developed algorithms require less human assistance. Frankel, Jennings and Lee (2016) apply support-vector regressions to analyze how words and bigrams from the MD&A section of 10-K reports explain firm-level accruals. The paper finds that these words and bigrams can predict accruals better than existing models and also help predict future cash flows. Also, Manela and Moreira (2017) use support vector regression (SVR) to derive an uncertainty measure called news implied volatility (NVIX) from Wall Street Journal lead stories (WSJ). They demonstrate a correlation between the occurrence of certain words in WSJ articles and the implied volatility of options (VIX) between 1996 and 2009. Specifically, they find that the greater the level of uncertainty expressed in the

WSJ stories (represented by a higher NVIX Index), the greater the following market stock returns. These findings also have economically significant implications since an increase of one standard deviation in NVIX is connected with a 3.3% rise in yearly returns the following year. The study conducted by Donovan et al. (2021) utilized three distinct machine learning techniques, namely support vector regressions, supervised Latent Dirichlet Allocation, and random forest regression trees, to develop a credit risk metric that draws on qualitative data obtained from conference calls and the MD&A section of 10-K reports. The study reveals that the measure enhances the accuracy of forecasting credit events and explains the fluctuations in credit risk both at the inter-firm and intra-firm levels.

Other studies applying supervised learning to textual analysis include Taddy (2013), Rabinovich and Blei (2014), and Taddy (2015). As opposed to the supervised learning techniques, Bybee et al. (2020) extend the research in this area by employing an unsupervised modeling technique called Latent Dirichlet Allocation (LDA) to evaluate the topics covered by the Wall Street Journal. In contrast to Manela and Moreira (2017), that only look at the most prominent WSJ stories, Bybee et al. (2020) analyze the whole body of WSJ articles published between 1984 and 2017. They find that there is a correlation between media coverage and economic growth. These studies all confirm the usefulness of machine learning in text analysis. However, one of the critical limitations of employing machine learning is the black-box issue. The operations of specific algorithms are overly sophisticated to be understood by humans due to their size and complexity. These complex algorithms can hardly produce a transparent and tractable model. Besides, machine learning methods sacrifice intuition and understanding for predictive accuracy and efficiency (Bzdok, Altman and Krzywinski, 2018). It remains unclear whether machine learning is understanding texts' logic or just extracting signals (Frankel, Jennings and Lee, 2021).

2.3 Textual analysis and social media

Content mined from Yahoo stock forums was among the first examples of textual analysis' application to the financial sector (see Das and Chen, 2007). However, in recent years, Twitter has become one of the most cutting-edge sources of social media information. Twitter, which launched as a microblogging service in 2006, has become one of the most popular platforms for individuals to convey or disseminate new information. Whether these short tweets could have an impact on the firm's fundamentals and

capital market has also been discussed by many researchers. Bollen, Mao and Zeng (2011) analyzed the mood expressed through Twitter. Their findings show that incorporating public sentiment characteristics can enhance the accuracy of DJIA forecasts. Mao et al. (2012) examines whether the daily number of tweets mentioning the S&P500 stocks is associated with S&P500 stock indicators (share price and traded volume). They argue that the movement of the S&P500 Index closing prices can be predicted more accurately when including Twitter data in the model. With the help of Twitter to distribute links to press releases and other conventional disclosures, Blankespoor, Miller and White (2014) investigates whether companies may utilize Twitter as an additional communication channel to spread information better. After controlling for the news's content, the existence of information intermediaries, market circumstances, and firm-specific features, they show that Twitter distribution during news event windows is linked with reduced bid-ask spreads, bigger depths, and a higher liquidity ratio, especially for those companies with low exposure to the traditional financial press. Another similar research by Lee, Hutton and Shu (2015) argues that businesses utilize Twitter and other social media to communicate with investors in an effort to mitigate the market's unfavorable reaction to the news. Moreover, according to a study by Jung et.al (2017), most S&P1500 businesses have created either a corporate Twitter account or a Facebook profile. More importantly, Jung et.al (2017) shows that a company's tweets about news and followers' retweets could trigger more news coverage from traditional media. It indicates that the news press and Twitter can simultaneously interact and affect the capital market. Azar and Lo (2016) analyze the predictive ability of social media to forecast the returns of a value-weighted stock index by creating a database of tweets about the Federal Reserve's Federal Open Market Committee (FOMC). They assign each tweet a polarity score between -1 and +1. The weighting of the tweets is based on the number of followers each individual has, as individuals with more followers should have a more significant influence on index results. They discovered that the investor tweeted opinions regarding the Federal Reserve affect the performance of value-weighted indexes. Higher Fed-related sentiment on Twitter is associated with greater future index returns, and the results are exceptionally high on the eight FOMC dates. Bartov, Faurel and Mohanram (2018) investigated tweets posted nine trading days prior to the announcement of earnings. They apply two alternative methodologies to quantify the sentiment of tweets. The first method classifies tweets as either good, negative, or neutral based on the results of a naive Bayes algorithm. The second method

employs three lexicons to evaluate negative sentiment: the negative word list by Loughran and McDonald (2011), the Harvard IV-TagNeg word list, and the word list developed by Hu and Liu (2004). Both methods have demonstrated their ability to forecast forthcoming actual earnings surprises and prompt stock price reactions to earnings declarations. Elliott, Grant and Hodge. (2018) observes that in an experimental scenario, the impact of CEO Twitter accounts mitigates the effect of unfavorable news. Following negative corporate news, investors are more inclined to purchase the stock of a company whose Chief Executive Officer engages with market on Twitter. Similarly, Agrawal, Gans and Goldfarb (2018) take both Twitter and StockTwits messages into account. The authors suggest that there exists a correlation between social media posts and liquidity indicators, specifically turnover and the quantity of intraday trades that occur beyond the quotation spread. Besides, there exists a correlation between raised bullish or bearish sentiment and increased liquidity metrics. Moreover, the authors posit that the influence of negative effect on liquidity surpasses that of positive effect; therefore, the authors contend that moments of panic are anticipated to have a more pronounced effect on market turnover and trading volume compared to optimistic sentiment. Other similar literature includes Liew and Budavári (2016), Nofer and Hinz (2015), and Porshnev, Lakshina and Redkin (2016).

More recently, Gu, Kelly and Xiu (2020) investigated the data content of firm-specific sentiment derived from Twitter, using Bloomberg's supervised machine learning algorithms for social media sentiment analysis. They argue that tweets contain information not reflected in stock prices. Gan et al. (2020) apply Thomson Reuters MarketPsych Indices (TRMI) to measure the sentiment, and they conclude that the social media, not news, dominated the causal association between media sentiment and market characteristics (return and volatility). Also, they have discovered that the correlation between implied volatility and sentiment exhibits more robustness in comparison to the correlation between returns and sentiment. Cookson and Niessner (2020) creatively measure investor disagreement by observing investors' investing model through Stocktwits message postings. By analyzing disagreement within and across investment strategies, they determine the extent to which different information sets versus different interpretations of information cause disagreement. Cao, Fang and Lei (2021) collect tweets from individual firms to measure Negative Peer Disclosure (NPD). Then the peer disclosure tweets are categorized according to the quantity of positive or negative lexicon as specified in the Loughran and McDonald dictionary. They find that the tendency to issue NPD rises with product

market competition and technology proximity. Except for the associations between Twitter information and essential market variables such as return, and volatility, it has also been observed that Twitter integrates financially essential information for bond and credit default swap investors who are assumed to possess specialist expertise in the field (Bartov, Faurel and Mohanram, 2022). By applying the dictionary-based text analysis, Bartov, Faurel and Mohanram (2022) construct their main variable aggregated Twitter opinion (OPI) from a factor analysis using three commonly used word lists: the Loughran and McDonald (2011) word list, the Harvard IV word list, and the Hu and Liu (2004) word list.

As immediate and worldwide information sources, Twitter and other social media platforms appear to be treasure troves of timely, globally relevant market information. However, the usage of slang, vulgarity, symbolism, and sarcasm is one of the challenges in gauging mood from social media posts. Loughran and McDonald's (2011) lexical standards might be used as a starting point for this kind of sentiment analysis. However, they were not developed to analyze the sentimental content of Twitter posts. Both Renault (2017) and Chen, Guo and Renault (2019) provide an example of how crucial it is to tailor sentiment-related vocabulary lists to the target corpus. Loughran and McDonald's dictionary, for instance, was compiled with 10-K reports. Researchers should modify the underlying vocabulary lists cautiously and transparently before applying the dictionary to tweets, volatility, or any other source.

2.4 Volatility modelling and forecasting

Researchers have established several models aimed at capturing the dynamics of market volatility. Engle (1982) proposed the Autoregressive Conditional Heteroskedastic (ARCH) model, which analyses stylized characteristics of price volatility, such as persistence, mean reversion, and heavy tails, by formulating conditional variance as a linear function of past observables. Bollerslev (1986) extended the original ARCH model to create the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, which has been extensively used and discussed by practitioners and researchers. Evolving from these theoretical cornerstones, several variations of the GARCH model have been proposed, focusing on different variables. For example, GARCH-M models (Engle, Lilien and Robins, 1987), EGARCH models (Nelson, 1991), and modified GARCH-M models (Glosten, Jagannathan and

Runkle, 1993). Although GARCH models can be easily applied, they perform poorly in forecasting from the high-frequency intraday data available for many financial assets. As Anderson et al. (2003) proposed, realized volatility—calculated based on high-frequency asset return data—is an unbiased estimator of return volatility. It is evident that the conventional volatility models that are employed for daily-level prediction, such as ARCH and GARCH, are not equipped to process the mass of information contained in intraday data. These models tend to fall short when used in attempts to accurately represent longer interdaily volatility fluctuations. Corsi's original (2009) HAR model, with simple parameters, significantly outperforms the GARCH and stochastic volatility models (Bollerslev, Patton and Quaedvlieg, 2016). Much of the literature is devoted to the construction of various modified versions of the HAR model, such as the HARQ model (Bollerslev, Patton and Quaedvlieg, 2016) and the vector HAR model (Busch, Christensen, and Nielsen, 2011). This dissertation focuses on the impact of words, so we consider HAR to be a satisfactory fundamental model to be integrated with word variables. We therefore developed PCA-augmented HAR models using daily Twitter text as data and performed forecasts to investigate whether words from the tweets predicted S&P500 Index return volatility. Audrino, Sigrist, and Ballinari (2020), also employed the HAR model, with the inclusion of sentiment and attention variables, to forecast future volatility. In their case, the sentiment and attention variables were obtained using a deep learning algorithm, utilizing a large volume of Stock Twits posts and Google searches as their basis.

2.5 Contributions

The present research fills a gap in the existing literature because there has hitherto been no investigation of how Twitter content can predict future asset return volatility. We link stock return volatility directly to individual words using the PCA method. For textual analysis, several techniques have been used to extract meaningful information from text data, with manual classification, bag-of-words, and machine learning being among the most common approaches. Each method has its own advantages and disadvantages. However, the practicality of the bag-of-words approach has made it a popular choice for many applications. Manual classification offers a high degree of accuracy and incorporates expert knowledge. However, manual classification is labor-intensive, time-consuming, and prone to human error or bias. Furthermore, it scales poorly with increasing data volumes, making it less suitable for

handling large datasets. Machine learning techniques, such as deep learning and natural language processing algorithms, are capable of handling more complex representations of text data. However, machine learning approaches have been criticized for picking up trends or signals without genuinely understanding the text content, which makes it impossible for researchers to confirm whether the algorithms are truly learning from the context or are simply identifying hidden features that are incomprehensible to human logic. The bag-of-words approach is a text data simplification technique that represents language as a collection of words, recording the occurrence frequency of individual words without considering grammar or word sequence. This approach offers several benefits, including ease of implementation, computational efficiency, and scalability to large datasets. However, in previous studies, the ‘bags-of-words’ were created by professionals, which raises a subjectivity issue. Besides, as the number of concepts increases, and are represented differently over time due to the evolution of language and economics, and as expertise on different topics or even the same topic increases and disseminates, a need arises for discipline and replicability in the construction of word bags.

The aim of this dissertation is to refine the process of constructing bags of words for the study of volatility. Our strategy first replaces subjective human decision-making with a reproducible algorithm (PCA). Since this technique is well-defined and straightforward, it can be replicated. Second, it can be modified and enhanced by future researchers, making it easier to evolve and adapt. Third, it provides experts from different fields with a baseline model from which to apply their knowledge to enhance their findings. This collaborative approach can lead to more comprehensive and accurate results. Finally, as opposed to machine learning algorithms that analyse long sentences or even the whole text of reports, focusing on the occurrences of single words avoids the "black box" issue and makes the entire process of generating word predictors more transparent.

Chapter 3. Data

Collecting Twitter text data related to the S&P500 Index involved using the Twitter Developer Platform. This platform provides access to the vast amount of real-time data generated by Twitter users. Our primary focus is on tweets directly connected to the S&P500 Index, necessitating a search for tweets containing a keyword closely related to the S&P500 Index. We could, of course, simply search for "S&P500," but this harvests many irrelevant tweets, for example:

S&P 500 is asserting a near-term trading range: Technical Indicator https://on.mktw.net/2MFVfcI
The 25 Best S&P 500 Stocks of the Past 50 Years https://nytv.to/Peyk
Did you know that more than half of the S&P 500 companies have agreed to be more transparent about their political spending and now share information that would otherwise be opaque and untraceable? Find out more in @rgrdlaw's #CorporateGovernance Roundup: https://bit.ly/2MYWbch
Weekly S&P500 ChartStorm WriteUp
Charting the markets with @IGTV@IGBank Key levels in #cable #euro #brentcrude and #S&P500 with @JeremyNaylor_IG Long term charts in Cable point to $1.36 key for medium term trend. Wishing you all and IGcredible team a happy and healthy 2022! #technicalanalysis #fx #charting
When will people realize that the real chair of the #FederalReserve is the S&P500?

These texts that mention "S&P500" contain a great deal of noise, and often include irrelevant or even misleading information. As an alternative, we chose to search for the keyword "SPX," which is the ticker symbol of the S&P500 Index. This more targeted approach sacrifices comprehensive coverage for relevance and reduced ambiguity. People who are knowledgeable about the stock market, such as traders and financial analysts, tend to use ticker symbols rather than full names when discussing stocks or indices. By searching for "SPX," we are more likely to find tweets from people with a deep understanding of the market, which improves the accuracy of our sentiment analysis. Moreover, unlike other potential keywords, "SPX" is less likely to be used in other contexts. For example, "S&P" could refer to the credit rating agency Standard & Poor's. Our approach enabled us to filter out irrelevant tweets and concentrate solely on data pertinent to our research. A similar approach was applied by Mao, Counts and Bollen (2011). Maintaining the integrity and relevance of collected data is also crucial. To ensure the quality and relevance of the data, we removed tweets created for promotional purposes, which could skew our data, impairing its accuracy for further analysis.

Twitter, the popular microblogging service launched in 2006, has experienced remarkable growth since its inception, from just 5000 daily users in 2007 to over 200 million worldwide by 2011. Such incredible

growth speaks to Twitter's immense popularity and its capacity to connect people around the globe. Our research aims to collect as much information about context and variation in tweets related to the S&P500 Index as possible. In order to accomplish this goal, we collected samples over a time range from 1 January 2012 through 31 December 2021 to ensure that we captured a wide range of variations in opinion and sentiment over time. However, it should be acknowledged that the negation problem is a potential limitation of this dissertation. Although the negation of positive sentiment is more common in 10-K reports than in tweets, it may obscure true sentiment and impact the validity of our analysis results.

Figure 1 details the pre-processing steps undertaken to prepare the data for analysis. Pre-processing passed the data through several essential stages, including punctuation removal, URL removal, stop-word removal, and tokenization of text data. Punctuation marks and URLs do not contribute significantly to Twitter user sentiment, so removing them is crucial for producing accurate results. Stop-word removal is essential for ensuring data quality. Stop-words are common words, for example ‘and’, ‘the’, and ‘is’, that do not provide any meaningful insights into Twitter users' sentiments and opinions about the S&P500 Index. Eliminating these stop words allowed us to focus on more crucial terms that provide greater value. Tokenization involves breaking sentences into individual words to accurately record frequency data for each term used, which helps in determining which terms appear most often. By implementing the above steps, we were able to generate time series data for each individual word, reflecting their occurrence on specific trading days. We also used the data to identify the top 150 high-frequency word series for further analysis. As Manning and Schütze (1999) posited, word counts tend to follow a power-law distribution. This implies that just a very few high-frequency words can substantially affect outcomes (Loughran and McDonald, 2016).

[Figure 1]

To collect realized variance data on the S&P500 Index, we utilized the Realized Library database developed by the Oxford-Man Institute of Quantitative Finance. This resource offers daily nonparametric measures of past index volatility and provided reliable and accurate measurements of realized variance for this vital component of our research. We were able to obtain intraday returns of the S&P500 Index at 5-minute frequency for the entire 2012-2021 sample period from which we generated daily realized variance figures with 5-minute frequency. As well as measuring realized

variance (5-minute), we also collected an alternative measure of realized volatility: realized kernel variance. This nonparametric estimator of integrated variance uses a smooth weighting kernel to reduce microstructure noise impact. This alternative measure of realized volatility was collected to provide a robust comparison to the realized variance, ensuring that our results were comprehensive and accurate.

Figures 2 and 3 illustrate the logarithmic realized variance series and the level realized variance series using two different measures. Specifically, Figure 3 presents the five most frequently occurring words during three instances of exceedingly high market volatility: August 24, 2015, February 6, 2018, and March 12, 2020. These dates correspond to three notable market crashes. On August 24, 2015, the S&P500 Index experienced a significant drop of 103.88 points within minutes compared to the closing value on August 21. On February 6, 2018, the S&P500 Index declined by 4.1%, closing at 2,648.94. Lastly, on March 12, 2020, the S&P500 Index reached 2,741.38, following a 4.9% decrease. An examination of Figure 3 reveals that "low" and "down" were frequently mentioned during these periods, signifying a prevalence of negative sentiment among investors. There are also examples of tweets during these days of the crash, where people express their astonishment and feelings of panic regarding the behavior of the market:

Date	*Text*
24/08/2015	*US equity futures are pointing to a bloody open: $DJIA -535, $SPX -56 with just under 2 hours until the open*
24/08/2015	*Global meltdown continues, $SPX futs have accelerated lower from last night, now -60, or 3%, on par w/ Europe's losses, China finished -8.5%*
24/08/2015	*Futures crashed to new lows overnight http://t.co/wNG47HwQH9 $SPY $SPX $DJIA $DIA $VIX $COMPQ*
06/02/2018	*An Unprecedented Move To Financial Crisis Lows https://t.co/22TdzfWKPQ #stocks #amrkets $DJIA $SPX*
06/02/2018	*US #StockMarket Futures Slumping Overnight After Initial Plunge $NDX $SPX https://t.co/urEOA00RpI*
06/02/2018	*The Dow is heading for another nosedive this morning, falling more than 300 points premarket $DJI $SPX https://t.co/23uDy2rE1b*
12/03/2020	*Market halted right out of open. $SPX $SPY $TSLA #optionstrading #stockstotrade https://t.co/skczUxiTBA*
12/03/2020	*Another #market wide 15min circuit breaker. $DJI $DJIA $SPX $COMP $IXIC*
12/03/2020	*hope you realize your witnessing history on legacy markets $SPX $DJI $SPY https://t.co/mvz5bT7AAG*

[Figure 2]

[Figure 3]

Table 1 shows that the mean values for realized variance (5-minute) and realized kernel variance are very close, ranging from 12.39% to 12.93%. The minimum values are also similar, ranging from 1.50% to 1.75%. However, the maximum values show some variation, with the realized variance (5-minute) having a higher value of 101.90%. The most frequently occurring word is "es_f," which stands for E-mini-S&P500 futures contracts, with a mean of 109 and a total of 274,297 occurrences. The least frequently occurring word is "sensex," which stands for Bombay Stock Exchange Sensitive Index, with a mean of 2 and 4,240 occurrences. The word with the highest standard deviation is "qqq," with a value of 98. "qqq" stands for the Invesco QQQ Trust Series 1, which is an ETF based on the Nasdaq-100 Index. The word with the highest maximum occurrence is "good," with a value of 1,059. Almost all words have a minimum daily occurrence of 0. These statistics provide insight into the popularity and usage patterns of various 'hot' words in tweets related to the S&P500 Index over the sample period. Figure 4 displays a 'word cloud' in which the size of each word is determined by the frequency with which it occurs. Words that appear more frequently are represented by larger font sizes.

[Table 1]

[Figure 4]

Chapter 4. Methodology

4.1 Realized variance.

The log-price process P of a single asset during an active trading day as it develops in continuous time can be described by the following equation:

$$dP_t = \mu_t dt + \sigma_t dW_t \tag{1}$$

Where μ and σ are the instantaneous drift and volatility processes, and W represents standard Brownian motion. The ith Δ-period return within day t is defined as

$$r_{i,t} = P_{t-1+i\Delta} - P_{t-1+(i-1)\Delta},\ i = 1, 2, \dots, M, \tag{2}$$

Where $M = 1/\Delta$ is the sampling frequency. Hence the daily logarithmic return for the active part of the trading day t is $r_t = \sum_{i=1}^{M} r_{i,t}$.

We normally forecast the latent one-day integrated variance defined by

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds \tag{3}$$

Although one-day integrated variance is not directly observable, it can be reliably inferred from the one-day realized variance (Anderson et al. (2003), Barndorff-Nielsen, Ole E and Shephard (2002) and Barndorff-Nielsen and Barndorff-Nielsen, Ole E and Shephard (2006)). It is defined as follows:

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2 \tag{4}$$

where $r_{t,i}$ is the i_{th} intraday return on day t.

4.2 The HAR-PCA model

The HAR model has the benefit of modeling the long-memory behavior of volatility in a straightforward

and parsimonious manner, resulting in excellent forecasting performance when applied to the measurement of realized volatility (Corsi, 2009). The original HAR model specifies realized variance as a linear function of daily, weekly, and monthly realized variance components:

$$RV_t = \beta_0 + \beta_1 RV_{t-1}^{d} + \beta_2 \overline{RV}_{t-1,\mathrm{t-5}}^{w} + \beta_3 \overline{RV}_{t-1,\mathrm{t-22}}^{m} + \varepsilon_t \quad (5)$$

Where $RV_{t-1}^{d} = RV_{t-1}$ is the daily realized volatility at $t-1$. $\overline{RV}_{t-1,\mathrm{t-5}}^{w}$ is the average realized volatility over the past week. $\overline{RV}_{t-1,\mathrm{t-22}}^{m}$ is the average realized volatility over the past month. The simplicity of the model also allows for augmentation with other statistically or economically significant regressors. The PCA-augmented model used in this study is defined as follows:

$$RV_t = \beta_0 + \beta_1 RV_{t-1}^{d} + \beta_2 \overline{RV}_{t-1,\mathrm{t-5}}^{w} + \beta_3 \overline{RV}_{t-1,\mathrm{t-22}}^{m} + \textstyle\sum_n \gamma_n * PCA_{t-1}^{n} + \varepsilon_t \quad (6)$$

Where PCA_{t-1}^{n} is lagged one day PCA variable that derives from the number of occurrences of the specific word during the most recent day. This specification captures the lagged effect of words in addition to the lagged realized volatility by applying the same structure as the HAR model.

4.3 Principal component analysis

PCA is a technique used in statistics to reduce the dimensionality of a dataset while retaining as much of its variance as possible. Stock and Watson (2002) first applied this method in an economic context. Their results suggest that PCA produces consistent and asymptotically efficient forecasts even under general assumptions about cross-sectional and temporal dependence among the variables. Ehsani and Linnainmaa (2022) also utilized PCA to extract factors that explain more of the cross-section of returns. In essence, PCA is a widely used unsupervised learning technique in the field of data analysis and machine learning. It is a dimensionality reduction method that transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components. These components capture the maximum variability in the original dataset while reducing noise and redundancy. PCA relies on a fundamental mathematical theorem called the singular value decomposition (SVD) theorem which

states that any symmetric matrix, including a variance-covariance matrix such as Σ, can be decomposed into its symmetric eigenvalues and eigenvectors (see discussion in Guidolin and Pedio, 2020).

$$\Sigma = U\Lambda U' \tag{7}$$

where U is a square $n \times n$ matrix, the ith column of which is the eigenvector u_i of Σ and Λ is a diagonal matrix, whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$. U is an orthogonal matrix, so that $UU' = U'U = I_n$. Using this decomposition, the principal components of the input data can be computed as linear combinations of the original variables. The vector of the ith principal component is given by

$$PCA_i = U'x_i \tag{8}$$

where x_i represents the vector of the original variables. PCA involves several steps, which include pre-processing, covariance matrix calculation, eigenvalue and eigenvector calculations, dimension reduction, and transformation of data.

Pre-processing: As part of pre-processing, data must first be standardized so that each feature has zero mean and one variance value to ensure that each feature contributes equally to covariance matrix calculations.

Covariance matrix calculation: Once the data have been standardized, the next step is to calculate their covariance matrix to gain information regarding the relationships among features within it. This step provides insights into any patterns or correlations that may exist within the data.

Eigenvalue and eigenvector calculation: The covariance matrix is decomposed into its eigenvalues and eigenvectors. The eigenvectors represent the directions along which the data have the largest variance, and the eigenvalues represent the amount of variance in each direction.

Dimensionality reduction: The eigenvectors corresponding to the largest eigenvalues are used to project the data onto a lower-dimensional subspace. The number of dimensions in the subspace can be controlled by specifying the number of components to keep during projection.

Transformation of the data: Finally, the data is transformed by projecting it onto the lower-dimensional subspace defined by the eigenvectors. The transformed data has reduced dimensionality while retaining as much as possible of the information contained in the original data.

We compiled the frequency of every word that appeared each day throughout the sampling period and generated time series for the top 150 words. We subsequently conducted PCA on these variables, following the steps described, to reduce the dimensionality of the words' variables. The selected principal components—for example, PCA1, PCA2, ..., PCAN—are the first, second, and Nth principal components, respectively. These components represent the directions along which the data have the greatest variance. The first principal component, PCA1, is the direction in which the data has the maximum variance. In other words, PCA1 captures the most crucial underlying pattern in the data. The second principal component, PCA2, is orthogonal to PCA1 and captures the next most important pattern in the data. The third principal component, PCA3, is orthogonal to both PCA1 and PCA2 and captures the third most important pattern in the data, and so on. To interpret these principal components, we look at their loadings on the original variables (word frequency variables). The loadings are obtained by regressing each component on all the word variables. Loadings indicate the strength and direction of the relationships between variables and components. Word variables with high absolute values (close to 1 or -1) on each principal component are the most important for defining each component and distinguishing it from others.

4.4 In-sample and out-of-sample forecasting framework

We used a portion of our time series data (01/01/2012-31/12/2013) to estimate the model parameters and determine the best-fitting model. This is identified as the in-sample data. We first used PCA to construct a set of independent variables that summarize the textual content, which we used to augment the HAR model for forecasting realized volatility. We then applied three information criteria: the Akaike Information Criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Schwartz, 1997), and the Hannan-Quinn Information Criterion (HQIC) (Hannan and Quinn, 1979) to select the model that performs best. These criteria are defined as follows:

$$\mathrm{AIC} = -2 * \log(\mathrm{L}) + 2 * \mathrm{k} \tag{9}$$

$$\mathrm{BIC} = -2 * \log(\mathrm{L}) + \mathrm{k} * \log(\mathrm{n}) \tag{10}$$

$$HQIC = -2 * log(L) + 2k * log(log(n)) \tag{11}$$

where $\log(\mathrm{L})$ is the logarithm of the likelihood function of the model, k is the number of parameters in the model, and n is the sample size. These information criteria are used to compare and select the best model among the candidates. The lower the value of an information criterion the better the fit of the model.

Once the model has been estimated using the in-sample data, it is important to assess its performance using previously unseen data. We therefore evaluated the forecasting model's performance using a separate observation from 01/01/2014 to 31/12/2021; the OOS period. The OOS period contains data not used during the model development stage. To make a one-day ahead prediction, we then use the available information up to the most recent time step and apply our forecasting methodology to predict next-day volatility. This process is repeated sequentially for each time step in the OOS set, updating the model's input with the newly observed data each time. The updating frequency is another hyperparameter that needs to be determined carefully. In this study, we applied three updating frequencies: daily, weekly, and monthly. With a one-day updating frequency, the forecasting model is updated, and new predictions are generated, on a daily basis. This means that, every day, the model incorporates the most recent available data to update its parameters and make forecasts for the next day's volatility. In the case of a one-week updating frequency, the forecasting model is updated once a week (five trading days) while new predictions are generated on a daily basis. The same logic applies to the monthly (22 trading days) updating frequency. We used three different updating frequencies, but always forecast one-day ahead volatility. In addition, we used both rolling- and expanding-window schemes, with the length of the rolling window being two years.

4.5 Forecasting performance metrics

Following Patton (2011), Patton and Sheppard (2009), and Greene (2003), the mean squared error (MSE), mean absolute percentage error (MAPE), and Theil's U were used to evaluate the OOS forecasting performance. The OOS R-squared value was also considered, following Campbell and

Thompson (2008). The original HAR model was estimated as a benchmark for assessing forecasting accuracy when word variables were entered. The MSE, MAPE, Theil's U ratio and OOS R-squared are defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(RV_i - \widehat{RV_i})^2 \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{RV_i - \widehat{RV_i}}{RV_i}\right| \tag{13}$$

$$Theil's\ U = \sqrt{\frac{\sum_{i=1}^{n-1}(\frac{\widehat{RV}_{i+1} - RV_{i+1}}{RV_i})^2}{\sum_{i=1}^{n-1}(\frac{RV_{i+1} - RV_i}{RV_i})^2}} \tag{14}$$

$$R^2_{OOS} = 1 - \frac{MSE_i}{MSE_{benchmark}} \tag{15}$$

where n is the number of forecasts, RV_i is the actual ith value, and $\widehat{RV_i}$ is the forecasted ith value. MSE_i is the mean squared error of model i, and $MSE_{benchmark}$ is the mean squared error of the benchmark model.

When comparing OOS forecasting performance, lower values for MSE, MAPE, and Theil's U ratio indicate better performance, while for OOS R-squared, a higher value suggests better forecasting performance.

Chapter 5. Empirical results

5.1 In-sample estimates

As discussed in part 4.3, Table 2 can be used to understand the main themes or topics captured by each principal component (PC). For example, on PC1, the words "qqq," "es_f," "ndx," and "market" have high positive loadings, which means that they are positively correlated with PC1 and explain most of its variation. PC1 therefore seems to be related to market indices and volatility. On the other hand, for PC2, the words "qqq," "es_f," and "aapl" are still highly correlated loadings, but the correlations are negative. As a result, "qqq" and "es_f" may be seen as ways to get exposure to the S&P500 Index. PC1 could be 'high attention to the index' while PC2 could be 'low attention to the index.' For PC3, the words "down," "sell," "close," and "low" have high positive loadings. This suggests that PC3 represents negative sentiment or bearish trends in the market. PC3 therefore seems to be related to market movements and support levels. PC4 and PC5 both seem to capture positive sentiment, with PC4 having a twist towards European markets. This is consistent with their negative loadings on volatility.

[Table 2]

As indicated in Table 3, by incorporating principal components as supplementary regressors, the log-HAR model is able to harness more information from high-frequency data, thus enhancing its predictive capabilities. A notable trend observed in the table is the general improvement in the log-HAR model's fit as more PCAs are incorporated, as demonstrated by the progressively higher R^2 values. However, it is important to note that not all PCAs exhibit statistical significance at any confidence level. Specifically, PCA1, PCA2, PCA6, PCA7, and PCA8 do not contribute significantly to the model. In contrast, PCA3, PCA4, and PCA5 emerge as the most significant PCAs, with PCA4 and PCA5 displaying negative signs. This suggests that these two PCAs are capturing negative correlation patterns in RV. Moreover, the table reveals that the coefficients of the monthly, weekly, and daily averages of logarithmic RV consistently remain positive and statistically significant across all models. This indicates their ability to capture long-term, medium-term, and short-term persistence in RV, thereby enhancing the model's overall predictive power. Lastly, the constant term is consistently negative and significant throughout all log-HAR models, highlighting the mean-reverting property of RV.

[Table 3]

Table 4 shows that the model with the lowest AIC value is log-HAR-PCA6, which means that it is the

model most preferred by AIC. The model with the lowest BIC value is log-HAR, and the model with the lowest HQIC value is log-HAR-PCA5. AIC, BIC, and HQIC have different properties and assumptions and may not always agree on the best model. For example, BIC and HQIC penalize more complex models compared to AIC. Table 4 indicates that there is no clear consensus on which forecasting model among the candidates is the best. The fact that a model exhibits good performance on in-sample data does not guarantee its ability to generalize to OOS data. It is therefore necessary to conduct additional evaluations to assess the OOS performance of each model more accurately.

[Table 4]

5.2 Out-of-sample results

5.2.1 Full sample period

Prior to delving into the forecasting results, Table 5 shows the correlation matrix of all independent variables employed in the HAR-PCA models. This reveals that lagged realized variance variables exhibit high correlations amongst themselves. In contrast, the PCs demonstrate negligible or no correlations with each other, which is a fundamental feature of the PCA method. Each principal component represents a different aspect of the information contained in the original data. Because they are orthogonal, these components are completely unrelated to each other. The most significant correlation between the PCs and the lagged realized variances is observed between PC3 and the one-day lagged realized variance, with a value of 0.3895. This indicates that these two components share some common information, but not enough to cause serious multicollinearity problems. It is also worth noting that the correlations between PC1 and the three lagged realized volatility variables are close to 0.3000, and the correlation between PC3 and the three lagged realized volatility variables ranges from 0.2172 to 0.3895.

[Table 5]

According to Table 6, HAR-PCA models generally demonstrate superior performance to the HAR model in terms of Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Theil's U, particularly when employing the Rolling Window (RW) estimation method. This implies that HAR-PCA models can extract more valuable information from word frequency data compared to the HAR model alone, suggesting that Twitter sentiment has a substantial impact on the dynamics of volatility.

The MSE metric assesses the average squared difference between realized volatility's actual and forecast values. The results indicate that HAR-PCA models typically have lower MSE values than the HAR model, especially when applying the RW estimation method. Among all PCA models, log-HAR-PCA-5 with the RW method achieves the lowest MSE (0.4121), improving the forecasting error of the original HAR model under RW (0.4225) by 2.46%. Moreover, log-HAR-PCA-5 with the RW method also boasts the lowest MAPE (0.0493), making it the most effective model for predicting future volatility according to the MAPE metric. The R-squared and Theil's U values corroborate the MSE and MAPE results, reinforcing the superior performance of log-HAR-PCA-5 under the RW method. In addition to outperforming the HAR model, the forecasts generated by log-HAR-PCA-5 using the RW method are, statistically, significantly different from those of the HAR model. This finding suggests that the principal components extracted from Twitter text data can enhance the volatility forecasting of the S&P500 Index when integrated with HAR model predictors, capturing more variation than the original HAR forecasting method. Notably, the table also reveals that incorporating more principal components does not consistently lead to improved forecasting performance. In some cases, PCA models with fewer components yield lower errors than models with a higher number of components. This suggests that the word frequency data may contain some redundancy or residual noise that the PCA models are unable to eliminate.

[Table 6]

Figure 5 plots the forecasting value of log-HAR-PCA-5 (RW) and log-HAR (RW) models compared with realized variances series. It is evident that both of our models are able to capture the trend of the realized variance effectively. The difference between the prediction values of the two models and the actual realized variance is minimal, indicating that the models have a high level of accuracy. However, it is not clear why the performance of the log-HAR-PCA-5 model is superior. Figure 6 shows that the log-HAR-PCA-5 model improves on the forecasting performance of the original log-HAR model by accurately capturing extreme changes in volatility. Specifically, the plot indicates that, when there are significant fluctuations in variance (either high or low), the log-HAR-PCA-5 model provides more accurate predictions than the log-HAR (RW) model, as evidenced by the higher or lower forecast values. This improvement can be attributed to the incorporation of Twitter word frequency data, which supplies additional information by reflecting social media sentiment. In addition to evaluating the predictive

performance of logarithmic realized variance, Figures 7 and 8 also depict performance in forecasting realized variance by computing the exponential of the logarithmic forecasted values. Notably, Figure 8 demonstrates that the log-HAR-PCA-5 model is more accurate in predicting when the realized variance attains exceptionally high levels, a finding that aligns with the observations in Figure 6.

[Figure 5]

[Figure 6]

[Figure 7]

[Figure 8]

The Social media sentiment, as gleaned from Twitter word frequencies, offers valuable insights into the public's perception and expectations regarding market conditions, which in turn influence market volatility. Integrating this additional information into the log-HAR-PCA-5 model endows it with higher explanatory power in forecasting extreme volatility in the S&P500 Index.

Table 7 shows that the HAR-PCA models also outperforms the HAR model when utilizing less frequent estimation. However, the degree of improvement is less than that observed in Table 6. The primary reason for this is the difference in updating frequency for parameter estimation between the two tables. In Table 7, the updating frequency is set at five days, as opposed to the daily updating frequency used in Table 6. As a consequence, the responsiveness and adaptability of the models to fluctuations and evolving patterns in the data might be somewhat diminished. A closer look at the HAR-PCA models reveals that log-HAR-PCA-5, employing the rolling window method, exhibits the best overall performance. With the lowest MSE of 0.4160, MAPE of 0.0495, and Theil's U of 0.9078, along with the highest R-squared value of 0.6836, log-HAR-PCA-5 demonstrates superior forecasting ability. Furthermore, these models are statistically distinct from the HAR model at various significance levels, thus reinforcing the merits of the PCA approach. Similarly, the HAR-PCA models with a greater number of components show higher error rates than those with fewer components.

[Table 7]

Table 8 shows that HAR-PCA models do not perform better than the HAR model in forecasting realized volatility when the updating frequency for parameter estimation is 22 days: i.e., on a monthly basis rather than daily or weekly. This may make the models less accurate and reliable, as they cannot capture the latest information and trends in the data. In fact, on this frequency cycle, some HAR-PCA models

have higher errors than the HAR model in MSE, MAPE, Theil's U and lower R-squared values. None of the HAR-PCA models significantly differ from the HAR model at any level. This suggests that word frequency data only provide valuable additional information for forecasting volatility when the updating interval is short. Among the HAR-PCA models, log-HAR-PCA-2 with the RW method has the lowest MSE (0.4219), MAPE (0.0497), Theil's U (0.9155), and the highest R-squared (0.6792). However, these models are not significantly better than the HAR model on any performance measure. The longer updating interval may negatively affect the accuracy and reliability of the models, as they become less capable of capturing the most recent information and trends within the data. Consequently, some HAR-PCA models exhibit higher error rates than the HAR model in terms of their MSE, MAPE, Theil's U, and lower R-squared values. In this scenario, none of the HAR-PCA models display statistically significant differences from the HAR model at any confidence level.

These results suggest that, when the updating interval is long, word frequency data does not provide additional information that supports more accurate forecasting of volatility. As implied by our earlier analysis, another potential reason for the varying performance of PCA models in forecasting realized volatility is that all of the principal components are derived from lagged one-day word frequencies. When the principal components are exclusively based on lagged one-day word frequency data, the models are inherently limited in their ability to capture longer-term patterns and trends in the financial markets. As the updating interval for parameter estimation increases, this limitation is likely to become more pronounced, leading to reduced forecasting performance. In this scenario, the HAR-PCA models may not be able to obtain relevant information from longer-term word frequency patterns, thereby rendering them less accurate and reliable in forecasting realized volatility.

[Table 8]

Overall, based on the results from Tables 6, 7, and 8, the HAR-PCA models tend to surpass the HAR model in terms of their ability to forecast future volatility. The updating frequency for parameter estimation is clearly a crucial factor that impacts the forecasting performance of HAR-PCA models. With shorter updating intervals (daily or weekly), the HAR-PCA models' forecasting performance is clearly better, with lower error rates (Mean Squared Error, Mean Absolute Percentage Error, and Theil's U) and higher R-squared values. In contrast, a monthly updating interval tends to impair the forecasting performance of HAR-PCA models, making them less accurate and reliable in capturing market trends

and patterns. It was also found that the number of principal components in HAR-PCA models does not consistently affect their forecasting performance. In some instances, HAR-PCA models with fewer components demonstrate superior performance, exhibiting lower error rates and higher R-squared values than models with more components. This finding highlights the importance of carefully selecting the optimal number of components for each specific application, as incorporating excessive number of components may introduce additional complexity and noise to the models without necessarily improving their forecasting capabilities.

5.2.2 Sub-sample period

In our analysis, we sorted the realized variance of each day in the OOS period in descending order. Then we divided the data into terciles based on their volatility levels: high, medium, and low volatility subsamples. We examined the forecasting performance of the models within each subsample to investigate each model's effectiveness. The primary objective of categorizing the full sample period into subsamples according to volatility level is to explore whether the forecasting performance of HAR-PCA models fluctuates with varying market conditions. Volatility levels indicate the degree of uncertainty, risk, and information asymmetry in the market. It is intriguing to analyze how HAR-PCA models perform under diverse volatility regimes.

Table 9 shows that HAR-PCA models generally outperform the HAR model in forecasting realized volatility across different levels of volatility but that the extent of improvement is not uniform. In the high and low volatility subsamples, the HAR-PCA models display lower MSE, MAPE, and Theil's U values than the HAR model, suggesting that they can forecast future volatility with greater accuracy and precision. On the other hand, for the medium volatility subsample, HAR-PCA models exhibit higher MSE values than the HAR model, implying that they are less capable of accurately forecasting future volatility than the HAR model. This discrepancy in forecasting performance emphasizes the importance of taking market conditions into consideration when employing PCA models for predicting volatility. The results imply that PCA models are more suitable for forecasting under conditions of particularly high or low volatility but are not suitable for medium conditions. It is also crucial to recognize that the OOS R-squared values for all the models—the HAR model and all HAR-PCA models—underperform

under medium volatility conditions. This signifies that such models are ill-suited for forecasting future volatility under medium volatility conditions, with their performances falling short even compared to the naïve model, which predicts future variance simply based on historical averages. This phenomenon can be attributed to the excessive concentration of sample variances at the medium level, ranging from -10.9648 to -10.0405 (log), as opposed to the low-level subsample's range of -13.6161 to -10.9648 (log) and the high-level subsample's range of -10.0382 to -5.4839 (log). The variability range of the medium-level subsample is therefore much smaller, so the average value forecasts generated by naïve models exhibit greater alignment with the realized variances.

To provide a more comprehensive explanation and comparison of the results in Table 9, we can look more closely at some specific examples of HAR-PCA models and examine how they outperform the HAR model. Take, for instance, the high volatility subsample. In this case, log-HAR-PCA-1 combined with the RW method yields an MSE of 0.5724, 3.31% lower than the corresponding HAR model with the RW method (0.5920). This demonstrates that the log-HAR-PCA-1 with the RW method can predict future volatility more accurately than the HAR model with the RW method. Moreover, the difference between the two models is statistically significant at 1% level, suggesting that this improvement in accuracy is not random. Similarly, when we examine log-HAR-PCA-2 combined with the EW method, we find that it achieves an MSE of 0.5588, which is 6.7% lower than the HAR model with the EW method (0.5989). This indicates that log-HAR-PCA-2 with the EW method can also forecast future volatility with greater precision than the HAR model with the EW method. Again, the difference between the two models is statistically significant at 1% level. The superior performance of log-HAR-PCA-1 and log-HAR-PCA-2 in high volatility environments can be attributed to their ability to capture the most important principal components, which account for most of the variations in word frequency data. The HAR-PCA models can effectively forecast volatility by focusing on these key components, outshining the HAR model in both the RW and EW.

Expanding on the analysis for low volatility subsamples reveals further distinctions between the HAR-PCA and HAR models in terms of their forecasting accuracy. For instance, log-HAR-PCA-8 with the RW method demonstrates an MSE of 0.3882, which is 5.69% lower than the HAR model with the RW method (0.4116). Similarly, log-HAR-PCA-7 with the EW method has an MSE of 0.3989, 8.76% lower than the HAR model with the EW method (0.4372). In both cases, the forecasts are statistically different

from those generated by the HAR model, at a significance level of 1%. This emphasises the PCA models' superior performance compared to the HAR models in the low volatility subsample. Other performance measures, such as the Mean Absolute Percentage Error (MAPE), Thiel's U, and R-squared values, produce results that are consistent with the MSE findings, further substantiating the improvement in the accuracy of the PCA models. The reason why log-HAR-PCA-8 performs exceptionally well in low volatility regimes can be attributed to its inclusion of more principal components. By incorporating more of these components, log-HAR-PCA-8 captures a larger proportion of the variations in word frequency data, thereby enhancing its ability to forecast volatility in low volatility environments. This ultimately leads to more accurate and reliable predictions compared to the HAR model.

These examples illustrate how varying numbers of principal components can yield differing results across distinct volatility regimes, depending on the extent to which they can extract information from word frequency data. In general, models with fewer principal components tend to perform more effectively in high volatility environments. Conversely, those with more principal components often excel in lower volatility settings (Tables 10 and 11). These characteristics mirror those found in Table 6.

[Table 9]

[Table 10]

[Table 11]

In In summary, the tables confirm that HAR-PCA models outshine HAR models when forecasting realized volatility across various levels of volatility, particularly when employing the RW method. The differences are most evident for high and low volatility subsamples; less so for medium subsamples. This implies that word frequency data has more significance and relevance for predicting volatility during periods characterized by extreme market uncertainty and risk. Further, the contrast between the RW and EW methods is also more pronounced for high volatility subsamples than for medium or low volatility subsamples, which suggests that the RW method is more adaptable and resilient to market conditions shifts than the EW method. The ideal number of principal components may depend on the volatility level and the parameter estimation method employed. There might therefore be trade-offs between capturing a higher degree of information from word frequency data and avoiding issues such as overfitting and noise. It is clearly essential to strike the right balance when selecting the most

appropriate principal components and estimation methods to maximize the HAR-PCA models' predictive accuracy.

Chapter 6. Robustness check

A robustness check is a statistical procedure that is used to assess the sensitivity of a research finding to changes in the data or analytical methods. In our case, we used an alternative measure of the dependent variable (realized kernel variance) that involved replacing the original dependent variable (realized variance (5-minute)) with a different measure that captured a related but distinct aspect. The check replicated the forecasting models, methodology, and explanatory variables used in the original analysis. If the two sets of results are consistent, the original process can be considered to be reliable.

6.1 Full sample period

The robustness check results in Table 12, Table 13 and Table 14 indicate that the HAR-PCA model's superior performance over the HAR model in forecasting accuracy remains consistent across both volatility measures. HAR-PCA models generally exhibit lower error rates (MSE and MAPE) and higher R-squared values than the HAR model, particularly when using the Rolling Window (RW) estimation method. The log-HAR-PCA-5 (RW) model again achieved the lowest MSE and Theils's U, and the highest OOS R-squared, while log-HAR-PCA-8 (EW) achieved the lowest MAPE. Moreover, the findings from the robustness check reveal that the choice of volatility measures used for forecasting does not significantly impact the sensitivity of the results to changes in the model specification and estimation methods. The statistical significance of the difference in forecasting performance between log-HAR-PCA-5 and the HAR model remains robust, suggesting that incorporating word frequency data as supplementary variables can enhance volatility forecasting for the S&P500 Index. Similar to the main analysis, the robustness check also demonstrates that incorporating more principal components does not consistently lead to improved forecasting performance, but that the choice of updating frequency for parameter estimation does significantly affect the forecasting performance of the HAR-PCA models. When using a shorter updating interval (daily or weekly), the forecasting performance of the HAR-PCA models is improved, characterized by lower error rates (MSE, MAPE, and Theil's U) and higher R-squared values. In contrast, a longer updating frequency tends to result in a decline in the forecasting performance of HAR-PCA models, with results similar to those of the original HAR model. Overall, the robustness checks support the study's main findings and demonstrate the robustness of the results to changes in the volatility measure used for forecasting. HAR-PCA models with word frequency

data as supplementary variables continually outperformed the HAR model in volatility forecasting, particularly when using shorter updating intervals for parameter estimation.

[Table 12]

[Table 13]

[Table 14]

6.2 Sub-sample period for robustness check

The results from the robustness check (Tables 15,16 and 17) indicate that the forecasting accuracy of the HAR-PCA models depends on the level of volatility, and on the updating frequency used. In general, HAR-PCA models outperform the HAR model in high and low volatility regimes but perform worse than the HAR model, and even than the naïve model, in medium volatility regimes. This can be attributed to the excessive concentration of sample variances at the medium level, which reduces the variation and information contained in the data. Furthermore, the results demonstrate that a shorter updating interval (daily or weekly) leads to better forecasting performance for HAR-PCA models, while a longer (monthly) updating interval leads to worse forecasting performance. The choice of the number of principal components included in the models may also depend on the level of volatility and the updating frequency used. The log-HAR-PCA-1 and log-HAR-PCA-8 models perform well in high and low volatility regimes, respectively. Overall, the robustness check confirms that HAR-PCA models have the potential to enhance the forecasting performance of HAR models when used under specific market conditions and when the appropriate level of volatility and updating frequency are selected.

[Table 15]

[Table 16]

[Table 17]

Chapter 7. Economic significance

Forecasting market volatility is a critical task for investors, as it helps them to anticipate potential risks and opportunities in their portfolios. Compared to the unconditional MV strategy, which assumes constant levels of risk and return, the conditional MV approach takes the dynamic nature of financial markets into account and adjusts investment allocations accordingly. The economic significance of this approach lies in its potential to improve portfolio performance by identifying and capitalizing on changes in market volatility. Investors can make informed decisions about asset allocation by employing statistical techniques to forecast volatility. Although some of our models' statistical results are better than those of the HAR model at a statistically significant level, this alone does not necessarily imply that our model has practical economic value. Therefore, to fully assess the real-world implications of our findings, we conducted an economic significance test. Economic significance testing involves evaluating the magnitude of an effect and determining whether it is large enough to be considered meaningful and relevant in the investment context. By conducting an economic significance test, following the process described by Taylor (2022), we can determine whether our models' results have practical implications and are worthy of consideration in decision-making processes.

In the study by Taylor (2022), trading rules and strategies are examined under three core assumptions. The first assumption posits that the conditional variance of asset prices influences excess returns, which can be modelled using a modified version of Merton's Intertemporal Capital Asset Pricing Model (ICAPM). The second assumption suggests that realized benefits stem from consuming excess returns and rely on the proportion of invested wealth and the user's risk preference. These benefits are encapsulated by the first two moments of returns. The third assumption introduces an active trading rule, known as the volatility timing strategy, which aims to maximize the unconditional expectation of realized benefits. This particular strategy is grounded in the first two moments of returns. Within the context of these trading strategies, the utility gain of a conditional MV strategy over an unconditional MV strategy is given by:

$$\mathcal{G} = \frac{(\exp[C^2]-1)\lambda_0^2}{4\theta\sigma^2} \tag{16}$$

In addition, the mean leverage can be calculated as:

$$\text{Leverage} = \frac{\mu+\lambda_0(\exp(C^2)-1)}{2\theta\sigma^2} \tag{17}$$

where λ_0 and θ are parameters that measure the market conditions and users' risk preferences, respectively, and C^2 is a measure of the forecaster's skill level. C represents the volatility of the expected logarithmic variance (vol-of-vol). μ denotes the excess return of the S&P500 Index over the yield of ten-year Treasury bonds, while σ^2 denotes the variance in excess returns. The utility gain is positive when $\lambda_0 > 0$ and $\exp[C^2] > 1$. The utility gain has three components: inverse risk preference ($1/\theta$), market conditions ($\lambda_0^2/4\sigma^2$), and the forecaster's skill level ($\exp[C^2] - 1$). The utility gain is independent of the level of noise in the stochastic variance measure. The result assumes a Gaussian distribution for the log of stochastic variance and its associated conditional expectation and error. However, the error term can have any distribution as long as it is independently distributed, and the unconditional expectation exists.

7.1 Empirical results

7.1.1 Full sample periods

Table 18 indicates that HAR-PCA models for each updating frequency (1, 5, or 22 days) generally outperform the HAR model, which only employs three lags of realized variance. This finding implies that HAR-PCA models can generate higher returns for investors who utilize them for forecasting volatility and adjust their portfolios accordingly. Furthermore, the rolling window (RW) method's forecasting value generally surpasses that of the expanding window (EW) method due to the former's superior forecasting skill. A key observation from Table 18 is that the log-HAR-PCA-8 (RW) model consistently demonstrates the highest utility gain among all HAR-PCA models for every updating frequency and market condition, with the exception of the longest updating interval (22 days), where the log-HAR-PCA-1 (RW) model marginally outperforms it. This finding suggests that incorporating eight principal components to forecast volatility delivers optimal economic performance across all the models examined in this study.

It is also important to note that the utility gain of both HAR-PCA and HAR models increases as λ_0 rises, emphasising the fact that better market conditions lead to higher returns for volatility forecasters. However, the utility gain is also contingent upon forecasting skill, which fluctuates across different HAR-PCA models and updating frequencies. For instance, when $\lambda_0 = \mu$ and the updating frequency = 1 day, the log-HAR-PCA-8 (RW) model boasts a utility gain of 0.1444 per annum, while the log-HAR model exhibits a utility gain of only 0.1321. The superior performance of the log-HAR-PCA-8 (RW) model, evidenced by its annual gain of 14.44%, can be attributed to the highly favourable market conditions prevailing during the out-of-sample period from 2014 to 2021. This is further confirmed by comparison with the historical mean annual return of the S&P500 Index over the same period, which was 13.24%. Another intriguing insight gleaned from Table 18 is that the utility gain of both HAR-PCA and HAR models declines as the updating frequency decreases since forecasting skill deteriorates when parameters are estimated less frequently.

[Table 18]

Table 19 sets out the mean leverage values required to obtain the utility gains in Table 18. The leverage values exceed 2 for all the models in Table 19, signifying a relatively higher level of debt for investors who employ the conditional MV strategy based on the log-HAR and log-HAR-PCA models. This elevated leverage implies that investors need to take on a considerable amount of debt to maximize their utility gains when utilizing these forecasting models in their decision-making process. Such a high level of leverage may not always be desirable, especially for risk-averse investors, or in uncertain market conditions, as it can lead to magnified losses during market downturns. Besides, according to Table 19, under a given risk preference (θ = 1), the leverage ratio increases as market conditions improve. This finding suggests that to achieve identical utility gains, investors must assume greater debt as market returns become less volatile. Furthermore, the table highlights that, for a fixed market condition, the leverage ratio fluctuates based on the forecasting model employed and the updating frequency. When all parameters in equation (17)—except for forecasting skill—remain constant, the sole variable influencing the change in mean leverage is forecasting skill itself. Consequently, mean leverage and forecasting skills follow a similar trend across forecasting models and updating frequencies. Superior forecasting skill necessitates a higher leverage ratio to reach a specific utility gain.

The log-HAR-PCA-8 RW model exhibits the highest leverage ratio. For this model, with an updating

frequency of 1 and a market condition of 0.0063, the leverage ratio stands at 2.4966. This implies that the investor must borrow $2.50 for every $1 of their own capital to achieve an annual utility gain of 0.9%. For the same model and updating frequency, but with a market condition of 0.1001, the leverage ratio rises to 4.4826, meaning that the investor must borrow $4.48 for every $1 of their own capital to achieve an annual utility gain of 14.44%. This further confirms that leverage ratios increase as market conditions improve. Intuitively, as market conditions become less volatile, investors seek additional leverage to boost their returns, which increases risk. Conversely, the log-HAR RW model with an updating frequency of 1 and a market condition of 0.0063 presents a mean leverage ratio of 2.4399, marginally lower than the log-HAR-PCA-8 RW model. This results in a lower annualized utility gain of 0.83%. When comparing the log-HAR RW model with the log-HAR-PCA-8 RW model, the latter improves the annualized utility gain by 8.43% per year while utilizing only an additional 2.32% leverage. In summary, the results suggest that investors need to adjust their leverage ratios based on market conditions to achieve the desired utility gains. As market conditions improve, investors must assume more debt to enhance their potential returns which, at the same time, increases their risk exposure. Notably, the results show that the log-HAR-PCA-8 RW model offers a higher annualized utility gain than the log-HAR RW model with only a marginal increase in leverage.

[Table19]

7.1.2 Subsample analysis

To further investigate subsample performance, we conducted an analysis of S&P500 Index excess returns and divided them into three subsamples based on the realized variance of the S&P500 Index. As detailed in 5.2.2, these subsamples were categorized as high, medium, and low volatility regimes. Extreme leverage is needed for the medium and low volatility regimes due to their low sample variance. To make the utility gains across three subsamples comparable, we calculated the value of θ such that the mean level of leverage was limited to 1, using μ and σ as values for the low volatility regime. In order to calculate θ, according to equation (13), we also incorporated the forecasting skill ($\exp[C^2] - 1$). Specifically, we computed the average forecasting skill for the HAR and eight HAR-PCA models in the low volatility regime and used this average value to calculate θ. Then we applied the value of θ thus obtained to calculate the utility gain for the high and medium volatility regimes.

Table 20 presents the economic performance of the eight HAR-PCA models and the HAR model in forecasting the realized variance of the S&P500 Index under the high volatility regime. The results suggest that investors can benefit from using HAR-PCA models to forecast volatility and adjust their leverage accordingly. The HAR-PCA models generally outperform the HAR model, consistent with the result during the all-sample period. The results also indicate that the log-HAR-PCA-8 (RW) model, which incorporates eight principal components to forecast volatility, consistently exhibits the highest utility gain among all HAR-PCA models for every updating frequency and market condition, except for the longest updating interval (22 days) under the high volatility regime. The log-HAR-PCA-5 (RW) model is the most favorable at a 22-day updating frequency. A comprehensive analysis can be conducted by comparing Table 20 with Tables 21 and 22. Table 21 demonstrates that, under the medium volatility regime, the log-HAR-PCA-1 (Rolling Window) model yields a higher annualized utility gain than either the log-HAR-PCA-8 (RW) or the HAR (RW) models when coefficients are re-estimated daily or monthly. Likewise, under the low volatility regime in Table 22, the log-HAR-PCA-1 (RW) model prevails across all updating frequencies. This finding suggests that incorporating solely the most critical feature extracted from word frequency can bolster the economic performance of HAR-PCA models under medium and low volatility regimes. For instance, when the updating frequency is set to 1, the log-HAR-PCA-1 (RW) model surpasses the HAR (RW) model by 2.52% during medium volatility periods and 3.22% during low volatility periods.

Overall, the economic performance of the HAR-PCA and HAR models is superior in low volatility regimes compared to medium and high volatility regimes. This is evidenced by the higher utility gains exhibited by HAR-PCA models and the HAR model in low volatility regimes for most scenarios, as opposed to those observed in medium and high volatility regimes. This outcome suggests that volatility timing strategies prove more effective in environments characterized by low market volatility rather than high or moderate market volatility.

[Table 20]

[Table 21]

[Table 22]

Table 23 shows that there is a clear relationship between mean leverage and forecasting skills, with the two following a similar trend. This is true at all scales and holds for the overall sample and for the

various subsamples. Additionally, by applying θ with the mean leverage of the low volatility subsample scaled to 1, we produce a risk preference of 0.0122, suggesting that investors in this subsample are more risk-averse than those across the full sample period. In the high volatility subsample (Table 23), the mean leverages for all forecasting models have negative values. This can be attributed to the fact that risk-averse investors tend to avoid borrowing money and investing in assets during times of market volatility. Instead, these investors are more inclined to lend money and earn interest, minimizing their exposure to potential risk. In the medium volatility subsample (Table 24), there is minimal difference between the log-HAR and log-HAR-PCA models. Consequently, the leverage ratios derived from these models also exhibit only marginal variations. Lastly, for the low volatility subsample shown in Table 25, the mean leverage values are close to 1. This suggests that, in this subsample, investors are more comfortable leveraging their investments, as lower volatility provides a more stable market environment.

In terms of the economic significance of volatility forecasting models, the findings reveal that HAR-PCA models generally surpass the HAR model in predicting the S&P500 Index's variances. The log-HAR-PCA-8 (Rolling Window) model demonstrates the greatest utility gain among all PCA models over the full sample period. The utility gain metric, which encompasses forecasting skill, risk preference, and market conditions, signifies that the utility gain of both HAR-PCA and HAR models escalates as market conditions ameliorate.

The subsample analysis further corroborates the superior economic performance of HAR-PCA models in forecasting volatility, with the log-HAR-PCA-8 (RW) model consistently exhibiting the highest utility gain in most scenarios within the high volatility regime. During medium and low volatility regimes, the log-HAR-PCA-1 (RW) model yields a greater annualized utility gain than either the log-HAR-PCA-8 (RW) or the HAR (RW) model when the coefficients are re-estimated daily or monthly. This observation suggests that incorporating only the most crucial feature extracted from word frequency data can enhance the economic performance of HAR-PCA models under medium and low volatility regimes. In summary, volatility timing strategies prove more effective under conditions of low market volatility as opposed to high or moderate volatility. Investors accordingly stand to gain from employing HAR-PCA models for forecasting volatility and adjusting their investment allocations, with the log-HAR-PCA-8 (RW) model consistently delivering optimal economic performance across most

scenarios.

[Table 23]

[Table 24]

[Table 25]

Chapter 8. Conclusions

The aim of this dissertation is to investigate how Twitter text data can be used to predict future asset return volatility. We collected a large sample of tweets related to the S&P500 Index from 2012 to 2021 and applied a PCA method to extract meaningful information from the text data. We then integrated the PCA-based word variables into a HAR model and performed volatility forecasts for the S&P500 Index. Our main findings are as follows: Firstly, Twitter text data contains valuable information that can improve volatility forecasting performance compared to the baseline HAR model. The PCA method is an effective and transparent way to construct word variables that capture the sentiment and opinions of Twitter users regarding the S&P500 Index. Moreover, the HAR-PCA models are more effective at forecasting under conditions of extremely high or low volatility. Secondly, our HAR-PCA model has higher economic value when applied to portfolio investment compared with the HAR model.

This study fills a research gap in textual analysis and volatility modelling by investigating the predictive power of Twitter texts for future asset return volatility. The research employs PCA to link stock return volatility directly to individual words and refines the process of constructing bags of words for volatility topics by employing PCA to replace subjective human decisions with a reproducible algorithm. This technique offers several benefits: it is well-defined and replicable, it can be enhanced by future researchers, it provides a baseline model for interdisciplinary collaboration, and it avoids the "black box" issue of machine-learning algorithms by empirically and transparently focusing on the occurrence of single words. More importantly, by capturing a more relevant informative representation of textual information, PCA enhances the simple bag-of-words method.

Our study has some limitations and provides some indications for the direction of future research. One limitation is that we searched for S&P500 Index related tweets using the keyword "SPX," which may have led to a less comprehensive collection of Twitter messages. Another limitation is that we focused on only one social media platform (Twitter) and one asset index (S&P500). Future research could extend our analysis method to other platforms and asset indices to test the robustness and generalizability of our results. The third limitation is that we use a 1-gram model to represent text data, which ignores grammar and word order. Future research could explore the application of more sophisticated natural language processing techniques that could capture more intricate text data features, for example, n-gram

models, Latent Dirichlet Allocation (LDA), and word embeddings.

While our research has economic significance for portfolio investment, it should be noted that transaction costs are ignored in our investment strategy. In real-world trading, these costs, which may include brokerage fees, bid-ask spreads, and taxes, can significantly impact the profitability of a strategy. Particularly in a strategy that involves frequent trading, these costs can accumulate over time and potentially outweigh the predicted gains, leading to an overestimation of the economic significance of our forecasts. This is acknowledged as a limitation of the research. For future research, we recommend incorporating an estimate of transaction costs into the analysis to provide a more accurate representation of potential returns.

Reference list

Agrawal, A., Gans, J. and Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp.716–723. doi: https://doi.org/10.1109/tac.1974.1100705.

Anderson, T.G., Bollerslev, T., Diebold, F.X. and Vega, C. (2003). Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American Economic Review*, 93(1), pp.38–62.

Antweiler, W. and Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), pp.1259–1294. doi:https://doi.org/10.1111/j.1540-6261.2004.00662.x.

Arif, S. and Lee, C. (2014). Aggregate investment and investor sentiment. *Review of Financial Studies*, 27(11), pp.3241–3279. doi:https://doi.org/10.1093/rfs/hhu054.

Audrino, F., Sigrist, F. and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, [online] 36(2), pp.334–357. doi: https://doi.org/10.1016/j.ijforecast.2019.05.010.

Azar, P.D. and Lo, A.W. (2016). The wisdom of twitter crowds: predicting stock market reactions to FOMC meetings via Twitter feeds. *The Journal of Portfolio Management*, [online] 42(5), pp.123–134. Available at: https://doi.org/10.3905/jpm.2016.42.5.123.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), pp.1645–1680. doi:https://doi.org/10.1111/j.1540-6261.2006.00885.x.

Baker, S.R., Bloom, N. and Davis, S.J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, [online] 131(4), pp.1593–1636. doi:https://doi.org/10.1093/qje/qjw024.

Barndorff-Nielsen, O. E and Shephard, N. (2002). Econometric analysis of realized volatility and its

use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(2), pp.253–280.

Barndorff-Nielsen, O. E and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of financial Econometrics*, 4(1), pp.1–30.

Bartov, E., Faurel, L. and Mohanram, P.S. (2022). The role of social media in the corporate bond market: Evidence from Twitter. *Management Science*, [online] pp.93–114. doi:https://doi.org/10.1287/mnsc.2022.4589.

Bartov, E., Faurel, L. and Mohanram, P.S. (2018). Can Twitter Help Predict Firm-Level Earnings and Stock Returns? *The Accounting Review*, 93(3), pp.25–57. doi:https://doi.org/10.2308/accr-51865.

Behrendt, S. and Schmidt, A. (2018). The twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *Journal of Banking and Finance*, [online] 96, pp.355–367. doi:https://doi.org/10.1016/j.jbankfin.2018.09.016.

Bianchi, D., Guidolin, M., and Pedio, M. (2020). *Dissecting time-varying risk exposures in cryptocurrency markets*. BAFFI CAREFIN, Centre for Applied Research on International Markets Banking Finance and Regulation, Università Bocconi.

Blankespoor, E., Miller, G.S. and White, H.D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. *The Accounting Review*, [online] 89(1), pp.79–112. doi:https://doi.org/10.2308/accr-50624.

Bochkay, K., Brown, S.V., Leone, A.J. and Tucker, J.W. (2022). Textual analysis in accounting: What's next? *Contemporary Accounting Research*. [online] doi:https://doi.org/10.1111/1911-3846.12825.

Bodnaruk, A., Loughran, T. and McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, [online] 50, pp.623–646. doi:https://doi.org/10.1017/s0022109015000411.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), pp.1–8.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), pp.307–327.

Bollerslev, T., Patton, A.J. and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), pp.1–18.

Botosan, C.A. (1997). Disclosure level and the cost of equity capital. *Accounting Review*, 72(3), pp.323–349.

Boudoukh, J., Feldman, R., Kogan, S. and Richardson, M. (2013). *Which news moves stock prices? A textual analysis*. National Bureau of Economic Research.

Bryan, S.H. (1997). Incremental information content of required disclosures contained in management discussion and analysis. *Accounting Review*, 72(2), pp.285–301.

Buccioli, A. and Kokholm, T. (2021). Shock waves and golden shores: the asymmetric interaction between gold prices and the stock market. *The European Journal of Finance*, 28(7), pp.1–18. doi:https://doi.org/10.1080/1351847x.2021.1897026.

Busch, T., Christensen, B.J. and Nielsen, M.Ø. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, 160(1), pp.48–57.

Bybee, L., Kelly, B.T., Manela, A. and Xiu, D. (2020). *The structure of economic news*. [online] Working paper of National Bureau of Economic Research. Available at: https://doi.org/10.3386/w26648.

Bybee, L., Kelly, B.T. and Su, Y. (2022). Narrative asset pricing: Interpretable systematic risk factors from news text. *SSRN Electronic Journal*. [online] doi:https://doi.org/10.2139/ssrn.3895277.

Bzdok, D., Altman, N. and Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(1), pp.5–6.

Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H. and Steele, L.B. (2013). The information content of

mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, [online] 19, pp.396–455. https://doi.org/10.1007/s11142-013-9258-3.

Campbell, J.Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4), pp.1509–1531.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2020). How to talk when a machine is listening?: Corporate disclosure in the age of AI. [online] Working paper of National Bureau of Economic Research. Available at: https://doi.org/10.3386/w27950.

Cao, S.S., Fang, V.W. and Lei, L. (2021). Negative peer disclosure. *Journal of Financial Economics*, [online] 140(3), pp.815–837. doi: https://doi.org/10.1016/j.jfineco.2021.02.007.

Chen, C., Li, G. and Renault, T. (2019). What makes cryptocurrencies special? investor sentiment and return predictability during the bubble. [online] Available at: https://doi.org/10.2139/ssrn.3398423.

Chung, S., Hung, C. and Yeh, C. (2012). When does investor sentiment predict stock returns? *Journal of Empirical Finance*, 19(2), pp.217–240.

Cookson, J. and Niessner, M. (2020). Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, [online] 75(1), pp.173–228. doi:https://doi.org/10.1111/jofi.12852.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), pp.174–196.

Cutler, D.M., Poterba, J.M. and Summers, L.H. (1990). Speculative dynamics and the role of feedback traders. *American Economic Review*, 80(2), pp.63–68.

Daniel, K., Hirshleifer, D. and Sun, L. (2019). Short-and long-horizon behavioral factors. *The Review of Financial Studies*, [online] 33(4), pp.1673–1736. doi:https://doi.org/10.1093/rfs/hhz069.

Das, S.R. and Chen, M.Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, [online] 53(9), pp.1375–1388. doi:https://doi.org/10.1287/mnsc.1070.0704.

Da, Z., Engelberg, J. and Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices.

The Review of Financial Studies, 28(1), pp.1–32.

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

Donovan, J., Jennings, J., Koharki, K. and Lee, J. (2021). Measuring credit risk using qualitative disclosure. *Review of Accounting Studies*, 26, pp.815–863.

Ehsani, S. and Linnainmaa, J.T. (2022). Factor momentum and the momentum factor. *The Journal of Finance*, 77(3), pp.1877–1919.

Elliott, W.B., Grant, S.M. and Hodge, F.D. (2018). Negative news and investor trust: The role of firm and CEO Twitter use. *Journal of Accounting Research*, [online] 56(5), pp.1483–1519. doi:https://doi.org/10.1111/1475-679x.12217.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica,* pp.987–1007.

Engle, R.F., Lilien, D.M. and Robins, R.P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, pp.391–407.

Firth, M., Wang, K. and Wong, S.M. (2015). Corporate transparency and the impact of investor sentiment on stock prices. *Management Science*, [online] 61(7), pp.1630–1647. doi:https://doi.org/10.1287/mnsc.2014.1911.

Frankel, R., Jennings, J. and Lee, J. (2016). Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, [online] 62(2-3), pp.209–227. doi:https://doi.org/10.1016/j.jacceco.2016.02.002.

Frankel, R., Jennings, J. and Lee, J. (2021). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, [online] 68(7), pp.5514–5532. doi:https://doi.org/10.1287/mnsc.2020.3911.

Freyberger, J., Neuhierl, A. and Weber, M. (2020). Dissecting characteristics nonparametrically. *The*

Review of Financial Studies, [online] 33(5), pp.2326–2377. doi:https://doi.org/10.1093/rfs/hhz123.

Gan, B., Alexeev, V., Bird, R. and Yeung, D. (2020). Sensitivity to sentiment: News vs social media. *International Review of Financial Analysis*, [online] 67, p.101390. doi:https://doi.org/10.1016/j.irfa.2019.101390.

Gao, Z., Ren, H. and Zhang, B. (2019). Googling investor sentiment around the world. *Journal of Financial and Quantitative Analysis*, 55(2), pp.549–580. doi:https://doi.org/10.1017/s0022109019000061.

Garcia, D., Hu, X. and Rohrer, M. (2022). The colour of finance words. *Journal of Financial Economics*, 147(3), pp.525–549. doi:https://doi.org/10.1016/j.jfineco.2022.11.006.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5), pp.1779–1801.

Greene, W.H. (2003). *Econometric analysis*. Pearson Education India.

Gu, C. and Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking and Finance*, 121, p.105969. doi:https://doi.org/10.1016/j.jbankfin.2020.105969.

Guidolin, M. and Pedio, M. (2021). Forecasting commodity futures returns with stepwise regressions: Do commodity-specific factors help? *Annals of Operations Research*, 299, pp.1317–1356.

Guyon, J. (2020). Inversion of convex ordering in the VIX market. *Quantitative Finance*, 20(10), pp.1597–1623. doi:https://doi.org/10.1080/14697688.2020.1753885.

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), pp.2333–2383. doi:https://doi.org/10.1093/rfs/hhaa009.

Hannan, E.J. and Quinn, B.G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), pp.190–195. doi: https://doi.org/10.1111/j.2517-6161.1979.tb01072.x.

Hansen, S., McMahon, M. and Prat, A. (2017). Transparency and deliberation within the FOMC: A

computational linguistics approach. *The Quarterly Journal of Economics*, 133(2), pp.801–870. doi:https://doi.org/10.1093/qje/qjx045.

Ho, K., Shi, Y. and Zhang, Z. (2013). How does news sentiment impact asset volatility? evidence from long memory and regime-switching approaches. *The North American Journal of Economics and Finance*, 26, pp.436–456. doi:https://doi.org/10.1016/j.najef.2013.02.015.

Hope, O. and Wang, J. (2018). Management deception, big-bath accounting, and information asymmetry: Evidence from linguistic analysis. *Accounting, Organizations and Society*, 70, pp.33–51. doi:https://doi.org/10.1016/j.aos.2018.02.004.

Huang, D., Jiang, F., Tu, J. and Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3), pp.791–837.

Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), pp.712–729.

Jiang, G.J. and Tian, Y.S. (2005). The Model-Free Implied Volatility and Its Information Content. *The Review of Financial Studies*, [online] 18(4), pp.1305–1342. doi:https://doi.org/10.1093/rfs/hhi027.

Jin, Z., Yang, Y. and Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 32, pp.9713–9729.

Jung, M.J., Naughton, J.P., Tahoun, A. and Wang, C. (2017). Do firms strategically disseminate? evidence from corporate use of social media. *The Accounting Review*, 93(4), pp.225–252.

Keynes, J.M. (1937). The general theory of employment. *The Quarterly Journal of Economics*, 51(2), pp.209–223.

Ke, Z., Kelly, B.T. and Xiu, D. (2019). Predicting returns with text data. *SSRN Electronic Journal*.

Kurov, A. (2010). Investor sentiment and the stock market's reaction to monetary policy. *Journal of Banking and Finance*, 34(1), pp.139–149.

Larcker, D.F. and Zakolyukina, A.A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2), pp.495–540. doi:https://doi.org/10.1111/j.1475-679x.2012.00450.x.

Law, K.K. and Mills, L.F. (2015). Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4), pp.777–819. doi:https://doi.org/10.1111/1475-679x.12081.

Lee, L.F., Hutton, A.P. and Shu, S. (2015). The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, 53(2), pp.367–404. doi:https://doi.org/10.1111/1475-679x.12074.

Liew, J.K. and Budavári, T. (2016). Do tweet sentiments still predict the stock market? *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.2820269.

Li, F. (2010). The information content of forward-looking statements in corporate filings-a naive bayesian machine learning approach. *Journal of Accounting Research*, 48(5), pp.1049–1102. doi:https://doi.org/10.1111/j.1475-679x.2010.00382.x.

Liu, S. (2023). Do investors and managers of active ETFs react to social media activities? *Finance Research Letters*, 51, p.103454. doi:https://doi.org/10.1016/j.frl.2022.103454.

Lopez-Lira, A. (2023). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35–65.

Loughran, T. and McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1), pp.1–11.

Loughran, T. and McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey.

Journal of Accounting Research, 54(4), pp.1187–1230. doi:https://doi.org/10.1111/1475-679x.12123.

Loughran, T., McDonald, B. and Pragidis, I. (2019). Assimilation of oil news into prices. *International Review of Financial Analysis*, 63, pp.105–118.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1), pp.137–162.

Manning, C., and Schutze, H. (1999). Foundations of statistical natural language processing. *MIT press.*

Mao, Y., Wei, W., Wang, B. and Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. In *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research* (pp. 69-72).

Mao, H., Counts, S., and Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.

McLean, D. and Zhao, M. (2014). The business cycle, investor sentiment, and costly external finance. *The Journal of Finance*, 69(3), pp.1377–1409.

Mian, G.M. and Sankaraguruswamy, S. (2012). Investor sentiment and stock market response to earnings news. *The Accounting Review*, 87(4), pp.1357–1384.

Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, pp.347–370.

Nofer, M. and Hinz, O. (2015). Using Twitter to predict the stock market. *Business and Information Systems Engineering*, 57, pp.229–242.

Oliveira, N., Cortez, P., Areal, N. (2013). On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume. In: Correia, L., Reis, L.P., Cascalho, J. (eds) Progress in Artificial Intelligence. EPIA 2013. Lecture Notes in Computer Science(), vol 8154. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40669-0_31

Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of*

Econometrics, 160(1), pp.246–256.

Patton, A.J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In: J.-P. Kreiß, T. Mikosch, R.A. Davis and T.G. Andersen, eds., *Handbook of Financial Time Series*. Springer, pp.801–838.

Polk, C. and Sapienza, P. (2004). The real effects of investor sentiment. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.585885.

Porshnev, A., Lakshina, V. and Redkin, I. (2016). Could emotional markers in Twitter posts add information to the stock market ARMAX-GARCH model? *Higher School of Economics Research Paper No. WP BRP*.

Purda, L. and Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3), pp.1193–1223.

Rabinovich, M. and Blei, D. (2014, January). The inverse regression topic model. In *International Conference on Machine Learning* (pp. 199-207). PMLR.

Rahimikia, E. and Poon, S.-H. (2020). Big data approach to realised volatility forecasting using HAR model augmented with limit order book and news. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.3684040.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking and Finance*, 84, pp.25–40. doi:https://doi.org/10.1016/j.jbankfin.2017.07.002.

Roll, R. (1988). R^2. *The Journal of Finance*, 43(3), pp.541–566. doi:https://doi.org/10.2307/2328183.

Schwartz, E.S. (1997). The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging. *The Journal of Finance*, [online] 52(3), pp.923–973. doi: https://doi.org/10.1111/j.1540-6261.1997.tb02721.x.

See-To, E.W.K and Yang, Y. (2017). Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27, pp.283–296. doi:https://doi.org/10.1007/s12525-017-0254-5.

Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P. (2017). Divergence of sentiment and stock market trading. *Journal of Banking and Finance*, 78, pp.130–141. doi:https://doi.org/10.1016/j.jbankfin.2017.02.005.

Smales, L.A. (2017). The importance of fear: investor sentiment and stock market returns. *Applied Economics*, 49(34), pp.3395–3421. doi:https://doi.org/10.1080/00036846.2016.1259754.

Soo, C.-K. (2018). Quantifying sentiment with news media across local housing markets. *The Review of Financial Studies*, 31(10), pp.3689–3719. doi:https://doi.org/10.1093/rfs/hhy036.

Stambaugh, R.F. and Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, 30(4), pp.1270–1315. doi:https://doi.org/10.1093/rfs/hhw107.

Stambaugh, R.F., Yu, J. and Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2), pp.288–302. doi:https://doi.org/10.1016/j.jfineco.2011.12.001.

Stock, J.H. and Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97, pp.1167–1179.

Sun, L., Najand, M. and Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking and Finance*, 73, pp.147–164.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108, pp.755–770. doi:https://doi.org/10.1080/01621459.2012.734168.

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3). doi:https://doi.org/10.1214/15-AOAS831.

Taylor, N. (2022). The determinants of volatility timing performance. *Journal of Financial Econometrics*, nbac002. doi:https://doi.org/10.1093/jjfinec/nbac002.

Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), pp.1139–1168. doi:https://doi.org/10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008). More than words: Quantifying language

to measure firms' fundamentals. *Journal of Finance*, 63(3), pp.1437–1467.

Tumarkin, R. and Whitelaw, R.F. (2001). News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3), pp.41–51. doi:https://doi.org/10.2469/faj.v57.n3.2449.

Van Binsbergen, J.H., Han, X. and Lopez-Lira, A. (2022). Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of Financial Studies*, hhac085. doi:https://doi.org/10.1093/rfs/hlab030.

Zhang, X., Fuehres, H. and Gloor, P.A. (2011). Predicting stock market indicators through twitter 'I hope it is not as bad as I fear'. *Procedia-Social and Behavioral Sciences*, 26, pp.55–62. doi:https://doi.org/10.1016/j.sbspro.2011.10.562.

Zhou, G. (2018). Measuring investor sentiment. *Annual Review of Financial Economics*, 10, pp.239–259. doi:https://doi.org/10.1146/annurev-financial-110217-022725.

Zouaoui, M., Nouyrigat, G. and Beer, F. (2011). How does investor sentiment affect stock market crises? Evidence from panel data. *Financial Review*, 46(4), pp.723–747. doi:https://doi.org/10.1111/j.1540-6288.2011.00318.x.

Tables and figures

Table 1. Summary statistics

Panel A

	Mean	Min	Max	Sample size
Realized volatility(5-min)	12.93%	1.75%	101.90%	2505
Realized kernel volatility	12.39%	1.50%	91.69%	2505

Panel B

Words	Mean	Std.	Max	Min	Total	Sample size
es_f	109	94	548	0	274,297	2505
qqq	101	98	628	0	252,094	2505
market	85	66	550	2	212,526	2505
vix	76	60	468	0	189,649	2505
iwm	55	51	441	0	137,594	2505
aapl	49	54	503	0	123,345	2505
down	46	51	510	0	115,785	2505
djia	43	33	313	0	108,596	2505
rut	37	40	347	0	93,540	2505
close	35	32	272	0	88,798	2505
futures	34	30	406	0	84,534	2505
like	33	32	352	0	81,667	2505
trade	32	35	552	0	80,787	2505
dia	32	31	328	0	79,657	2505
short	27	24	325	0	67,833	2505
low	25	28	319	0	63,727	2505
long	24	23	189	0	60,884	2505
good	24	30	1059	0	60,375	2505
support	24	26	491	0	59,278	2505
stockmarket	21	29	448	0	53,035	2505
highs	20	23	317	0	51,335	2505
gld	20	17	339	0	50,734	2505
dax	20	23	439	0	49,504	2505
buy	20	24	202	0	49,201	2505
top	19	25	438	0	48,067	2505
sell	19	28	432	0	47,988	2505
levels	18	18	174	0	45,738	2505
rally	18	20	249	0	44,875	2505
bullish	18	16	172	0	44,580	2505
lower	18	17	168	0	44,050	2505
resistance	17	18	206	0	43,816	2505
gold	17	17	259	0	43,469	2505
target	15	18	317	0	38,491	2505
range	12	12	175	0	29,566	2505
video	12	12	216	0	29,275	2505
forex	11	13	125	0	28,058	2505

bulls	11	11	88	0	27,130	2505
ftse	9	17	437	0	23,475	2505
uvxy	9	17	440	0	23,206	2505
eurusd	9	7	87	0	21,806	2505
min	8	10	235	0	21,110	2505
nyse	8	18	443	0	19,431	2505
study	8	6	62	0	19,251	2505
mkt	7	8	108	0	17,974	2505
crash	6	17	428	0	15,871	2505
euro	6	11	101	0	15,638	2505
cac	6	31	871	0	14,645	2505
cboe	5	16	534	0	11,534	2505
ibex	2	15	433	0	5,792	2505
sensex	2	15	435	0	4,240	2505

Panel A displays the summary statistics for the S&P500 Index's daily realized volatility(annualized), as measured by two different methods: realized volatility (5-minute) and realized kernel volatility. The sample size for each measure is 2505, and the table reports the mean, minimum, maximum, and sample size for each measure. Panel B presents the summary statistics for the daily occurrences of some hot words in tweets related to the S&P500 Index. The table lists 50 hot words sorted by their total occurrence during the sample period from January 1, 2012, to December 31, 2021. The sample size for each word is 2505, and the table reports each word's mean, standard deviation, maximum, minimum, and total occurrence.

Table 2. Loadings of principle components

PC1	Loadings	PC2	Loadings	PC3	Loadings	PC4	Loadings	PC5	Loadings	PC6	Loadings	PC7	Loadings	PC8	Loadings
qqq	0.4354	market	0.2237	down	0.3922	cac	0.4322	options	0.6812	great	0.2880	great	0.4923	dia	0.3083
es_f	0.3427	vix	0.2184	market	0.2371	stockmarket	0.2896	market	0.4232	good	0.2697	good	0.4466	qqq	0.3069
ndx	0.3012	ndx	0.1774	sell	0.1951	ftse	0.2280	top	0.3043	down	0.2607	earnings	0.2129	earnings	0.2629
market	0.2610	es	0.1223	close	0.1825	dax	0.2222	aapl	0.2801	support	0.2488	trade	0.2120	es	0.2251
vix	0.2322	stockmarket	0.1198	low	0.1705	vix	0.2178	free	0.0748	vix	0.2113	strong	0.2102	djia	0.1639
es	0.2200	options	0.1160	cac	0.1534	ibex	0.2166	marketwatch	0.0470	qqq	0.1989	support	0.1310	down	0.1561
aapl	0.2069	cac	0.1133	es_f	0.1420	sensex	0.2061	fed	0.0355	options	0.1286	best	0.1207	move	0.1355
iwm	0.2055	dax	0.0827	bounce	0.1273	nyse	0.2015	es_f	0.0298	correction	0.1178	sell	0.1203	market	0.1149
down	0.1606	dia	0.0815	support	0.1193	uvxy	0.1931	move	0.0290	best	0.1110	buy	0.1173	great	0.0998
rut	0.1560	earnings	0.0709	nyse	0.1130	ppprophet	0.1810	buy	0.0283	strong	0.1110	free	0.1159	aapl	0.0974
ibex	-0.0012	qqq	-0.0355	top	-0.0488	bear	-0.0670	iwm	-0.0425	buy	-0.0999	top	-0.0569	cboe	-0.0624
ppprophet	-0.0017	levels	-0.0358	update	-0.0602	lower	-0.0726	great	-0.0503	cac	-0.1085	bounce	-0.0586	price	-0.0685
marketwatch	-0.0019	zerohedge	-0.0359	trade	-0.0635	like	-0.0859	dax	-0.0522	djia	-0.1102	vix	-0.0784	money	-0.0863
ichimoku	-0.0022	key	-0.0402	earnings	-0.0654	es	-0.0913	cac	-0.0559	sell	-0.1326	rut	-0.0803	volume	-0.1013
fxe	-0.0023	intraday	-0.0464	es	-0.1161	close	-0.1012	good	-0.0602	earnings	-0.1515	low	-0.0822	highs	-0.1079
cboe	-0.0024	aapl	-0.0766	aapl	-0.1700	low	-0.1156	stockmarket	-0.0727	fed	-0.1528	ndx	-0.0940	top	-0.1209
forex	-0.0026	free	-0.1422	ndx	-0.1910	djia	-0.1246	rut	-0.1132	es	-0.1812	djia	-0.1177	options	-0.1515
stocktwits	-0.0039	daytrading	-0.1471	options	-0.2132	iwm	-0.1536	ndx	-0.1686	ndx	-0.2016	qqq	-0.1646	vix	-0.3585
zerohedge	-0.0054	dji	-0.2268	rut	-0.2308	ndx	-0.1589	qqq	-0.2078	dia	-0.2359	aapl	-0.2259	ndx	-0.3632
euro	-0.0076	es_f	-0.7766	qqq	-0.4099	down	-0.2104	vix	-0.2177	iwm	-0.4623	down	-0.3272	trade	-0.3882

The table analyzes the loadings of each word on the principal components (PCs). Loadings are coefficients that connect each original variable to a specific principal component and indicate the extent of each variable's contribution to the principal component. A high loading signifies a strong influence of the variable on the principal component, whereas a low loading denotes a weaker influence. Loadings can be either positive or negative, which reveals whether there exists a positive or negative correlation between the variable and the principal component. For instance, a positive loading for the word "market" on PC1 implies that an increase in "market" is accompanied by an increase in principal component 1. Conversely, a negative loading for "forex" on PC2 suggests that an increase in "forex" results in a decrease in principal component 2. The table can be split into two sections, where the top half of the table displays the top 10 positive loadings and the bottom half displays the top 10 negative loadings for each principal component.

Table 3. In-sample results

	log-HAR	log-HAR-PCA1	log-HAR-PCA2	log-HAR-PCA3	log-HAR-PCA4	log-HAR-PCA5	log-HAR-PCA6	log-HAR-PCA7	log-HAR-PCA8
$\overline{RV}^{m}_{t-1,t-22}$	0.2159**	0.2155**	0.2327**	0.1909**	0.1862*	0.2109**	0.2154**	0.2107**	0.2172**
	(0.0943)	(0.0944)	(0.0963)	(0.0968)	(0.0967)	(0.0961)	(0.0960)	(0.0982)	(0.1004)
$\overline{RV}^{w}_{t-1,t-5}$	0.2484***	0.2486***	0.2525***	0.2291***	0.2177***	0.2038**	0.1921**	0.1954**	0.1954**
	(0.0808)	(0.0809)	(0.0810)	(0.0809)	(0.0810)	(0.0804)	(0.0806)	(0.0820)	(0.0820)
RV^{d}_{t-1}	0.2902***	0.2905***	0.2789***	0.2611***	0.2615***	0.2456***	0.2297***	0.2293***	0.2292***
	(0.0512)	(0.0513)	(0.0529)	(0.0529)	(0.0528)	(0.0525)	(0.0535)	(0.0536)	(0.0536)
PCA1		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
PCA2			0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006
			(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
PCA3				0.0030***	0.0031***	0.0032***	0.0034***	0.0034***	0.0034***
				(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)
PCA4					-0.0020	-0.0020*	-0.0021***	-0.0021***	-0.0021***
					(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
PCA5						-0.0048***	-0.0049***	-0.0049***	-0.0049***
						(0.0015)	(0.0015)	(0.0015)	(0.0015)
PCA6							0.0026	0.0026	0.0026
							(0.0017)	(0.0017)	(0.0017)
PCA7								-0.0004	-0.0004
								(0.0019)	(0.0019)
PCA8									0.0007
									(0.0021)
Constant	-2.6191***	-2.6177***	-2.5246***	-3.3614***	-3.5179***	-3.5752***	-3.8116***	-3.8295***	-3.7657***
	(0.7346)	(0.7353)	(0.7426)	(0.7960)	(0.8005)	(0.7930)	(0.8069)	(0.8112)	(0.8363)
R^2	0.2597	0.2598	0.2610	0.2725	0.2763	0.2916	0.2949	0.2950	0.2952

The table provides an analysis of in-sample estimates for eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. The table outlines the coefficient estimates, standard errors (displayed within parentheses), and R^2 values for each log-HAR model. The dependent variable under examination is the logarithmic realized variance at time t, whereas the explanatory variables encompass the monthly, weekly, and daily averages of logarithmic realized variance, as well as the PCAs themselves.

Table 4. Model selection criteria (in-sample)

Criterion	log-HAR	log-HAR-PCA1	log-HAR-PCA2	log-HAR-PCA3	log-HAR-PCA4	log-HAR-PCA5	log-HAR-PCA6	log-HAR-PCA7	log-HAR-PCA8
AIC	1052.8879	1054.8644	1056.0370	1050.1704	1049.4969	1040.7998	**1040.4185**	1042.3623	1044.2582
HQIC	1059.5083	1063.1398	1065.9675	1061.7561	1062.7377	**1055.6956**	1056.9694	1060.5682	1064.1192
BIC	**-354.8519**	-348.6569	-343.2657	-344.9136	-341.3685	-345.8471	-342.0098	-335.8474	-329.7329

This table provides a comparison of the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and HQIC (Hannan-Quinn Information Criterion) values for all our forecasting models. The models are evaluated based on their goodness of fit and complexity, allowing for an informed selection of the most suitable model.

Table 5. Correlation matrix

	$\overline{RV}^{m}_{t-1,t-22}$	$\overline{RV}^{w}_{t-1,t-5}$	RV^{d}_{t-1}	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
$\overline{RV}^{m}_{t-1,t-22}$	1.0000										
$\overline{RV}^{w}_{t-1,t-5}$	0.8271	1.0000									
RV^{d}_{t-1}	0.6947	0.8505	1.0000								
PC1	0.2822	0.2966	0.3241	1.0000							
PC2	0.0111	0.0137	0.0108	0.0003	1.0000						
PC3	0.2172	0.3271	0.3895	-0.0003	-0.0002	1.0000					
PC4	-0.0636	-0.0768	-0.1028	0.0003	0.0002	-0.0001	1.0000				
PC5	0.0361	0.0632	0.0141	0.0003	0.0002	-0.0002	0.0001	1.0000			
PC6	0.1132	0.2091	0.2316	0.0005	0.0003	-0.0003	0.0003	0.0003	1.0000		
PC7	-0.0667	-0.1012	-0.1451	0.0002	0.0001	-0.0001	0.0001	0.0001	0.0002	1.0000	
PC8	0.1914	0.1325	0.1317	-0.0004	-0.0002	0.0002	-0.0002	-0.0002	-0.0004	-0.0001	1.0000

The table exhibits the correlation among every independent variable employed in the HAR-PCA models throughout the entirety of the sample period (01/01/2012-31/12/2021). The dependent variable is the realized variance of S&P 500 Index, and the independent variables include the lagged realized volatility at daily, weekly, and monthly frequencies, as well as the top eight principal components extracted from daily word frequency data.

Table 6. Out-of-sample performance for forecasting daily realized volatility (5-min) from 2014 to 2021.

Updating frequency=1		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.4225	0.4198	0.4166**	0.4143***	0.4146**	**0.4121*****	0.4152**	0.4147**	0.4132***
	EW	0.4239	0.4218	0.4222	0.4191	0.4158*	0.4144**	**0.4143***	0.4160	0.4178
MAPE	RW	0.0508	0.0499	0.0494	0.0494*	0.0495	**0.0493****	0.0494*	0.0495*	0.0494*
	EW	0.0511	0.0503	0.0504	0.0501	0.0498*	**0.0493****	0.0494**	0.0494*	0.0495
Theil's U	RW	0.9162	0.9122	0.9080	0.9055	0.9060	**0.9034**	0.9084	0.9084	0.9072
	EW	0.9201	0.9154	0.9157	0.9132	0.9072	**0.9046**	0.9047	0.9074	0.9101
R^2	RW	0.6787	0.6807	0.6832	0.6849	0.6847	**0.6866**	0.6843	0.6846	0.6858
	EW	0.6689	0.6706	0.6702	0.6726	0.6752	0.6763	**0.6764**	0.6750	0.6737
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analyses (PCA) models and the HAR models. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, was the benchmark model for computing out-of-sample R-squared values. Updating frequency=1 indicates the parameters of models are re-estimated on a daily basis

Table 7. Out-of-sample performance with weekly updating frequency.

Updating frequency=5		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.4235	0.4205	0.4183*	0.4174**	0.4181	**0.4160****	0.4227	0.4231	0.4218
	EW	0.4247	0.4226	0.4232	0.4220	**0.4193**	0.4200	0.4217	0.4244	0.4271
MAPE	RW	0.0509	0.0499	0.0495*	0.0495*	0.0496	**0.0495***	0.0498	0.0498	0.0498
	EW	0.0512	0.0504	0.0504	0.0502	0.0500	**0.0497**	0.0499	0.0499	0.0500
Theil's U	RW	0.9178	0.9130	0.9098	0.9087	0.9098	**0.9078**	0.9173	0.9180	0.9165
	EW	0.9212	0.9165	0.9169	0.9164	**0.9108**	0.9128	0.9151	0.9189	0.9239
R^2	RW	0.6779	0.6802	0.6819	0.6826	0.6820	**0.6836**	0.6786	0.6782	0.6792
	EW	0.6683	0.6699	0.6694	0.6704	**0.6725**	0.6720	0.6706	0.6685	0.6664
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, is the benchmark model for computing out-of-sample R-squared values. Updating frequency=5 indicates that the parameters of models are re-estimated on a weekly basis.

Table 8. Out-of-sample performance with monthly updating frequency.

Updating frequency=22		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.4263	0.4222	**0.4219**	0.4229	0.4233	0.4228	0.4280	0.4286	0.4284
	EW	0.4261	**0.4238**	0.4244	0.4305	0.4284	0.4309	0.4367	0.4394	0.4418
MAPE	RW	0.0512	0.0500	**0.0497**	0.0498	0.0498	0.0499	0.0502	0.0502	0.0502
	EW	0.0513	0.0504	0.0505	0.0506	0.0504	**0.0503**	0.0506	0.0506	0.0506
Theil's U	RW	0.9233	0.9158	**0.9155**	0.9165	0.9167	0.9165	0.9234	0.9243	0.9249
	EW	0.9238	**0.9183**	0.9189	0.9251	0.9202	0.9255	0.9317	0.9354	0.9370
R^2	RW	0.6758	0.6789	**0.6792**	0.6784	0.6781	0.6785	0.6745	0.6741	0.6742
	EW	0.6672	**0.6690**	0.6685	0.6638	0.6654	0.6634	0.6589	0.6568	0.6549
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, is the benchmark model for computing out-of-sample R-squared values. Updating frequency=22 indicates the parameters of models are re-estimated on a monthly basis.

Table 9. Sub-out-of-sample performance

Updating frequency=1		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.5920	**0.5724*****	0.5912	0.5881	0.5842	0.5842	0.5960	0.5917	0.5883
	EW	0.5989	0.5605***	**0.5588*****	0.5699***	0.5672***	0.5843	0.5809	0.5975	0.5999
MAPE	RW	0.0684	**0.0670*****	0.0682	0.0682	0.0679	0.0680	0.0686	0.0683	0.0680
	EW	0.0694	0.0665***	**0.0664*****	0.0670***	0.0667***	0.0678	0.0676	0.0684	0.0683
Theil's U	RW	0.9682	**0.9516**	0.9645	0.9620	0.9597	0.9598	0.9714	0.9693	0.9681
	EW	0.9808	0.9475	**0.9462**	0.9544	0.9490	0.9594	0.9573	0.9708	0.9743
R^2	RW	0.7469	**0.7554**	0.7473	0.7486	0.7503	0.7503	0.7452	0.7471	0.7485
	EW	0.7258	0.7434	**0.7441**	0.7391	0.7403	0.7325	0.7340	0.7264	0.7253
Subsample volatility level: Medium										
MSE	RW	0.2631	0.2697***	0.2604***	**0.2542*****	0.2563***	0.2553***	0.2579***	0.2615***	0.2625***
	EW	**0.2347**	0.2477***	0.2492***	0.2514***	0.2404*	0.2439***	0.2498***	0.2511***	0.2514***
MAPE	RW	0.0390	0.0395**	0.0389	0.0384*	0.0385	**0.0383***	0.0386	0.0388	0.0390
	EW	**0.0370**	0.0382***	0.0384***	0.0386***	0.0376**	0.0376	0.0381***	0.0382**	0.0383***
Theil's U	RW	0.7691	0.7809	0.7625	**0.7550**	0.7585	0.7567	0.7599	0.7656	0.7659
	EW	**0.7275**	0.7533	0.7555	0.7570	0.7384	0.7390	0.7489	0.7498	0.7506
R^2	RW	-0.4400	-0.4765	-0.4253	**-0.3917**	-0.4027	-0.3975	-0.4117	-0.4313	-0.4370
	EW	**-1.4940**	-1.6326	-1.6484	-1.6714	-1.5548	-1.5919	-1.6546	-1.6684	-1.6723
Subsample volatility level: Low										
MSE	RW	0.4116	0.4165*	0.3973***	0.3998**	0.4026	0.3961***	0.3912***	0.3905***	**0.3882*****
	EW	0.4372	0.4562***	0.4575***	0.4353	0.4390	0.4145***	0.4117***	**0.3989*****	0.4015***
MAPE	RW	0.0450	0.0454**	0.0444***	0.0446	0.0447	0.0443*	0.0440***	0.0441**	**0.0439*****
	EW	0.0468	0.0480***	0.0480***	0.0468	0.0471	0.0455***	0.0453***	**0.0447*****	0.0447***
Theil's U	RW	0.9635	0.9689	0.9453	0.9472	0.9500	0.9420	0.9359	0.9350	**0.9324**
	EW	0.9894	1.0109	1.0123	0.9877	0.9908	0.9621	0.9583	**0.9421**	0.9452
R^2	RW	0.7114	0.7080	0.7215	0.7197	0.7177	0.7223	0.7258	0.7262	**0.7278**

	EW	0.7208	0.7087	0.7079	0.7220	0.7197	0.7353	0.7371	**0.7452**	0.7436
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) and HAR models. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=1 indicates that the parameters of models are re-estimated on a daily basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility.

Table 10. Out-of-sample performance with weekly updating frequency.

Updating frequency=5		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.5950	**0.5741*****	0.5942	0.5920	0.5880	0.5875	0.6011	0.6010	0.5963
	EW	0.6006	0.5620***	**0.5610*****	0.5726***	0.5695***	0.5916	0.5898	0.6090	0.6142
MAPE	RW	0.0687	**0.0671*****	0.0683	0.0683	0.0680	0.0682	0.0687	0.0687	0.0686
	EW	0.0696	**0.0666*****	0.0666***	0.0672***	0.0668***	0.0685	0.0683	0.0693	0.0694
Theil's U	RW	0.9715	**0.9532**	0.9670	0.9650	0.9630	0.9633	0.9770	0.9778	0.9746
	EW	0.9827	0.9492	**0.9484**	0.9571	0.9512	0.9702	0.9694	0.9852	0.9935
R^2	RW	0.7457	**0.7546**	0.7460	0.7469	0.7486	0.7489	0.7430	0.7431	0.7451
	EW	0.7250	0.7427	**0.7431**	0.7378	0.7392	0.7291	0.7300	0.7212	0.7188
Subsample volatility level: Medium										
MSE	RW	0.2621	0.2690**	0.2595	**0.2546**	0.2574	0.2559	0.2636	0.2649	0.2670
	EW	**0.2342**	0.2472***	0.2486***	0.2526***	0.2402*	0.2434**	0.2510***	0.2513***	0.2509***
MAPE	RW	0.0389	0.0395**	0.0388	0.0383	0.0385	**0.0383**	0.0386	0.0387	0.0389
	EW	**0.0370**	0.0382***	0.0383***	0.0386***	0.0377**	0.0376*	0.0382***	0.0381**	0.0383**
Theil's U	RW	0.7677	0.7798	0.7612	**0.7567**	0.7616	0.7590	0.7721	0.7739	0.7760
	EW	**0.7268**	0.7526	0.7546	0.7595	0.7391	0.7393	0.7525	0.7518	0.7510
R^2	RW	-0.4349	-0.4722	-0.4204	**-0.3935**	-0.4089	-0.4010	-0.4426	-0.4500	-0.4615
	EW	**-1.4888**	-1.6276	-1.6420	-1.6842	-1.5534	-1.5872	-1.6672	-1.6714	-1.6669
Subsample volatility level: Low										
MSE	RW	0.4125	0.4175*	**0.4002*****	0.4046*	0.4082	0.4040*	0.4027*	0.4029*	0.4015*
	EW	0.4384	0.4577***	0.4589***	0.4400	0.4474**	0.4242***	0.4239**	**0.4124*****	0.4156***
MAPE	RW	0.0451	0.0455*	**0.0445****	0.0448	0.0450	0.0448	0.0447	0.0448	0.0446
	EW	0.0469	0.0481***	0.0481***	0.0470	0.0475**	0.0460***	0.0459***	**0.0453*****	0.0454***
Theil's U	RW	0.9644	0.9698	0.9483	0.9521	0.9555	0.9500	0.9482	0.9480	**0.9466**
	EW	0.9905	1.0124	1.0137	0.9925	0.9994	0.9722	0.9713	**0.9566**	0.9603
R^2	RW	0.7108	0.7073	**0.7194**	0.7163	0.7138	0.7168	0.7176	0.7175	0.7185

	EW	0.7200	0.7077	0.7069	0.7190	0.7143	0.7291	0.7293	**0.7366**	0.7346
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) and HAR models. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=5 indicates the parameters of models are re-estimated on a weekly basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility.

Table 11. Out-of-sample performance with monthly updating frequency.

Updating frequency=22		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.6024	**0.5779*** **	0.6012	0.6002	0.5974	0.5963	0.6021	0.5985	0.5971
	EW	0.6037	**0.5636*** **	0.5641***	0.5758***	0.5693***	0.5957	0.5967	0.6169	0.6179
MAPE	RW	0.0696	**0.0674*** **	0.0689	0.0689	0.0687	0.0688	0.0691	0.0688	0.0687
	EW	0.0699	**0.0668*** **	0.0669***	0.0675***	0.0669***	0.0688	0.0689	0.0701	0.0698
Theil's U	RW	0.9822	**0.9580**	0.9764	0.9755	0.9736	0.9737	0.9808	0.9791	0.9797
	EW	0.9873	**0.9517**	0.9521	0.9612	0.9526	0.9778	0.9795	0.9957	0.9956
R^2	RW	0.7425	**0.7530**	0.7430	0.7434	0.7446	0.7451	0.7426	0.7442	0.7448
	EW	0.7236	**0.7419**	0.7417	0.7363	0.7393	0.7272	0.7268	0.7176	0.7171
Subsample volatility level: Medium										
MSE	RW	0.2607	0.2688***	0.2560	**0.2522**	0.2527	0.2534	0.2595	0.2640	0.2647
	EW	**0.2330**	0.2464***	0.2466***	0.2506***	0.2427**	0.2458***	0.2545***	0.2545***	0.2539***
MAPE	RW	0.0388	0.0395***	0.0386	0.0382*	0.0381	**0.0381**	0.0385	0.0388	0.0388
	EW	**0.0369**	0.0381***	0.0381***	0.0383***	0.0378***	0.0377**	0.0384***	0.0383***	0.0384***
Theil's U	RW	0.7657	0.7796	0.7562	**0.7532**	0.7544	0.7550	0.7648	0.7718	0.7722
	EW	**0.7251**	0.7516	0.7519	0.7572	0.7441	0.7439	0.7584	0.7571	0.7556
R^2	RW	-0.4271	-0.4711	-0.4014	**-0.3803**	-0.3834	-0.3871	-0.4206	-0.4453	-0.4487
	EW	**-1.4761**	-1.6187	-1.6212	-1.6635	-1.5796	-1.6126	-1.7045	-1.7049	-1.6981
Subsample volatility level: Low										
MSE	RW	0.4149	0.4191	**0.4074**	0.4154	0.4189	0.4180	0.4217	0.4227	0.4229
	EW	**0.4407**	0.4603***	0.4614***	0.4640**	0.4724***	0.4505	0.4584	0.4464	0.4532
MAPE	RW	0.0452	0.0456	**0.0449**	0.0453	0.0455	0.0454	0.0457	0.0458	0.0457
	EW	0.0470	0.0482***	0.0483***	0.0477***	0.0483***	0.0469	0.0470	**0.0464***	0.0465
Theil's U	RW	0.9668	0.9716	**0.9567**	0.9646	0.9679	0.9665	0.9706	0.9713	0.9719
	EW	**0.9930**	1.0152	1.0164	1.0187	1.0264	1.0012	1.0090	0.9942	1.0014
R^2	RW	0.7091	0.7062	**0.7144**	0.7088	0.7063	0.7069	0.7043	0.7037	0.7035

	EW	**0.7185**	0.7060	0.7053	0.7037	0.6983	0.7123	0.7073	0.7150	0.7106
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) and HAR models. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=22 indicates the parameters of models are re-estimated on a monthly basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility.

Table 12. Robustness check: Out-of-sample performance for forecasting daily realized kernel from 2014 to 2021.

Updating frequency=1		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.5536	0.5502	0.5444**	0.5404***	0.5407***	**0.5366*** **	0.5410***	0.5393***	0.5370***
	EW	0.5547	0.5523	0.5530	0.5482	0.5436*	0.5400**	**0.5382** **	0.5397**	0.5421
MAPE	RW	0.0576	0.0567	0.0559**	0.0559**	0.0559**	0.0557***	0.0558**	0.0557***	**0.0556*** **
	EW	0.0578	0.0570	0.0570	0.0567	0.0562*	0.0556**	0.0557**	0.0555**	**0.0555** **
Theil's U	RW	0.8979	0.8927	0.8872	0.8833	0.8835	**0.8803**	0.8859	0.8849	0.8827
	EW	0.9026	0.8970	0.8975	0.8947	0.8878	**0.8829**	0.8818	0.8845	0.8874
R^2	RW	0.6092	0.6116	0.6157	0.6185	0.6183	**0.6212**	0.6181	0.6192	0.6209
	EW	0.5946	0.5963	0.5958	0.5993	0.6027	0.6053	**0.6066**	0.6055	0.6037
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, is the benchmark model for computing out-of-sample R-squared values. Updating frequency=1 indicates that the parameters of models are re-estimated on a daily basis. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 13. Robustness check: Out-of-sample performance with weekly updating frequency.

Updating frequency=5		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.5552	0.5515	0.5472**	0.5436***	0.5451**	**0.5427*** **	0.5506	0.5509	0.5488
	EW	0.5558	0.5536	0.5543	0.5514	0.5470*	0.5466*	**0.5463***	0.5494	0.5533
MAPE	RW	0.0577	0.0568	0.0561**	0.0561*	0.0562	**0.0561** **	0.0563	0.0563	0.0563
	EW	0.0579	0.0571	0.0571	0.0569	0.0564*	0.0561**	0.0562*	**0.0561***	0.0561
Theil's U	RW	0.9003	0.8944	0.8900	0.8864	0.8880	**0.8863**	0.8954	0.8958	0.8941
	EW	0.9039	0.8985	0.8990	0.8975	**0.8905**	0.8920	0.8920	0.8959	0.9024
R^2	RW	0.6081	0.6107	0.6137	0.6163	0.6152	**0.6169**	0.6113	0.6111	0.6125
	EW	0.5938	0.5954	0.5948	0.5970	0.6002	0.6005	**0.6007**	0.5985	0.5956
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, is the benchmark model for computing out-of-sample R-squared values. Updating frequency=5 indicates the parameters of models are re-estimated on a weekly basis. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 14. Robustness check: Out-of-sample performance with monthly updating frequency.

Updating frequency=22		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
MSE	RW	0.5591	0.5547	0.5520	**0.5512**	0.5527	0.5522	0.5586	0.5584	0.5588
	EW	0.5580	0.5556	0.5566	0.5587	**0.5555**	0.5572	0.5629	0.5662	0.5691
MAPE	RW	0.0581	0.0570	**0.0565**	0.0566	0.0567	0.0568	0.0570	0.0571	0.0570
	EW	0.0580	0.0572	0.0573	0.0572	0.0569	**0.0567**	0.0570	0.0569	0.0570
Theil's U	RW	0.9061	0.8987	0.8968	**0.8957**	0.8966	0.8968	0.9034	0.9035	0.9052
	EW	0.9069	0.9008	0.9017	0.9037	**0.8973**	0.9018	0.9062	0.9103	0.9118
R^2	RW	0.6053	0.6084	0.6103	**0.6109**	0.6098	0.6102	0.6056	0.6058	0.6055
	EW	0.5922	0.5939	0.5932	0.5916	**0.5940**	0.5927	0.5886	0.5862	0.5841
Sample size		2004	2004	2004	2004	2004	2004	2004	2004	2004

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both of the top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc. RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. The naïve forecasting model, which estimates future variance by relying on historical averages, is the benchmark model for computing out-of-sample R-squared values. Updating frequency=22 indicates the parameters of models are re-estimated on a monthly basis. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 15. Robustness check: Sub-out-of-sample performance with daily updating frequency

Updating frequency=1		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.7171	**0.6867*****	0.7162	0.7092	0.7008	0.6999	0.7173	0.7097	0.7039
	EW	0.7272	**0.6767*****	0.6766***	0.6911***	0.6863***	0.7114	0.7058	0.7322	0.7412
MAPE	RW	0.0758	**0.0739****	0.0753	0.0751	0.0745	0.0745	0.0752	0.0747	0.0744
	EW	0.0771	0.0739***	0.0738***	0.0746***	**0.0737*****	0.0751	0.0749	0.0761	0.0761
Theil's U	RW	0.9843	**0.9591**	0.9768	0.9706	0.9658	0.9653	0.9805	0.9761	0.9723
	EW	1.0027	**0.9626**	0.9628	0.9720	0.9635	0.9754	0.9718	0.9915	0.9988
R^2	RW	0.7027	**0.7153**	0.7030	0.7059	0.7094	0.7098	0.7026	0.7057	0.7081
	EW	0.6815	0.7036	**0.7036**	0.6973	0.6994	0.6884	0.6908	0.6793	0.6753
Subsample volatility level: Medium										
MSE	RW	0.3421	0.3505***	0.3347***	**0.3270*****	0.3278***	0.3271***	0.3303***	0.3345***	0.3373***
	EW	**0.3018**	0.3219***	0.3239***	0.3295***	0.3157*	0.3174	0.3251**	0.3199*	0.3191*
MAPE	RW	0.0437	0.0441*	0.0431**	0.0428***	0.0428***	**0.0428*****	0.0430*	0.0433	0.0436
	EW	**0.0412**	0.0423***	0.0425***	0.0429***	0.0420	0.0421	0.0428***	0.0426**	0.0426**
Theil's U	RW	0.7352	0.7474	0.7247	**0.7182**	0.7200	0.7186	0.7219	0.7266	0.7290
	EW	**0.6973**	0.7200	0.7219	0.7266	0.7088	0.7051	0.7153	0.7080	0.7079
R^2	RW	-0.8144	-0.8585	-0.7748	**-0.7339**	-0.7382	-0.7348	-0.7515	-0.7739	-0.7887
	EW	**-2.1988**	-2.4126	-2.4330	-2.4929	-2.3463	-2.3646	-2.4464	-2.3905	-2.3822
Subsample volatility level: Low										
MSE	RW	0.6013	0.6132***	0.5820***	0.5849***	0.5933	0.5827***	0.5754***	0.5737***	**0.5696*****
	EW	0.6292	0.6581***	0.6582***	0.6237	0.6286	0.5911***	0.5837***	0.5672***	**0.5662*****
MAPE	RW	0.0532	0.0541***	0.0525***	0.0527*	0.0530	0.0524**	0.0521**	0.0520**	**0.0516*****
	EW	0.0547	0.0563***	0.0563***	0.0546	0.0551	0.0529***	0.0524***	0.0514***	**0.0512*****
Theil's U	RW	0.9197	0.9286	0.9043	0.9054	0.9113	0.9030	0.8969	0.8961	**0.8926**
	EW	0.9360	0.9576	0.9578	0.9330	0.9361	0.9072	0.9009	0.8867	**0.8857**
R^2	RW	0.6365	0.6293	0.6481	0.6464	0.6413	0.6477	0.6521	0.6531	**0.6556**

	EW	0.6368	0.6202	0.6201	0.6400	0.6372	0.6588	0.6631	0.6726	**0.6732**
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc.RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=1 indicates that the parameters of models are re-estimated on a daily basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 16. Robustness check: Out-of-sample performance with weekly updating frequency.

Updating frequency=5		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.7223	**0.6904*****	0.7217	0.7159	0.7104	0.7083	0.7297	0.7311	0.7261
	EW	0.7298	**0.6794*****	0.6796***	0.6947***	0.6899***	0.7227	0.7174	0.7473	0.7607
MAPE	RW	0.0763	**0.0742****	0.0758	0.0756	0.0753	0.0751	0.0761	0.0761	0.0759
	EW	0.0774	0.0741***	0.0741***	0.0750***	**0.0740*****	0.0762	0.0760	0.0774	0.0776
Theil's U	RW	0.9901	**0.9633**	0.9821	0.9768	0.9749	0.9740	0.9922	0.9944	0.9913
	EW	1.0054	**0.9656**	0.9660	0.9758	0.9673	0.9923	0.9891	1.0111	1.0261
R^2	RW	0.7005	**0.7137**	0.7007	0.7032	0.7054	0.7063	0.6974	0.6968	0.6989
	EW	0.6803	**0.7024**	0.7023	0.6957	0.6978	0.6834	0.6857	0.6726	0.6668
Subsample volatility level: Medium										
MSE	RW	0.3408	0.3494**	0.3327**	**0.3239*****	0.3258***	0.3284**	0.3334	0.3327	0.3353
	EW	**0.3065**	0.3210***	0.3227***	0.3284***	0.3123	0.3135	0.3205**	0.3139	0.3118
MAPE	RW	0.0436	0.0441*	0.0430**	**0.0426*****	0.0427***	0.0428**	0.0429	0.0430	0.0433
	EW	**0.0414**	0.0423***	0.0425***	0.0428***	0.0418	0.0418	0.0426**	0.0422	0.0422
Theil's U	RW	0.7340	0.7465	0.7227	**0.7149**	0.7180	0.7209	0.7283	0.7263	0.7286
	EW	**0.6970**	0.7191	0.7207	0.7256	0.7049	0.7008	0.7105	0.7014	0.6994
R^2	RW	-0.8072	-0.8528	-0.7644	**-0.7174**	-0.7276	-0.7416	-0.7679	-0.7644	-0.7780
	EW	**-2.2491**	-2.4028	-2.4202	-2.4811	-2.3103	-2.3229	-2.3978	-2.3277	-2.3056
Subsample volatility level: Low										
MSE	RW	0.6023	0.6145***	0.5868***	0.5907**	0.5989	0.5914**	0.5888**	0.5889*	**0.5851****
	EW	0.6308	0.6602***	0.6604***	0.6308	0.6387	0.6036***	0.6009**	**0.5869*****	0.5874***
MAPE	RW	0.0533	0.0541***	0.0527**	0.0530	0.0533	0.0530	0.0529	0.0529	**0.0524**
	EW	0.0548	0.0564***	0.0564***	0.0549	0.0555***	0.0534***	0.0531***	0.0522***	**0.0521*****
Theil's U	RW	0.9203	0.9293	0.9072	0.9090	0.9146	0.9082	0.9056	0.9055	**0.9024**
	EW	0.9370	0.9588	0.9591	0.9374	0.9424	0.9153	0.9127	**0.9002**	0.9003
R^2	RW	0.6359	0.6285	0.6453	0.6429	0.6379	0.6425	0.6440	0.6440	**0.6463**

	EW	0.6359	0.6189	0.6188	0.6359	0.6313	0.6516	0.6531	**0.6613**	0.6610
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc.RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=5 indicates the parameters of models are re-estimated on a weekly basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 17. Robustness check: Out-of-sample performance with monthly updating frequency.

Updating frequency=22		log-HAR	log-HAR-PCA-1	log-HAR-PCA-2	log-HAR-PCA-3	log-HAR-PCA-4	log-HAR-PCA-5	log-HAR-PCA-6	log-HAR-PCA-7	log-HAR-PCA-8
Subsample volatility level: High										
MSE	RW	0.7320	**0.6977*****	0.7330	0.7312	0.7295	0.7258	0.7389	0.7324	0.7348
	EW	0.7344	**0.6825*****	0.6839***	0.6993***	0.6879***	0.7263	0.7267	0.7582	0.7632
MAPE	RW	0.0772	**0.0748*****	0.0768	0.0767	0.0767	0.0766	0.0773	0.0771	0.0773
	EW	0.0778	0.0744***	0.0745***	0.0754***	**0.0740*****	0.0767	0.0768	0.0784	0.0786
Theil's U	RW	1.0022	**0.9717**	0.9958	0.9941	0.9930	0.9923	1.0038	1.0008	1.0055
	EW	1.0112	0.9695	0.9708	0.9810	**0.9680**	0.9997	1.0002	1.0233	1.0255
R^2	RW	0.6965	**0.7107**	0.6960	0.6968	0.6975	0.6991	0.6936	0.6963	0.6953
	EW	0.6783	**0.7010**	0.7004	0.6937	0.6987	0.6818	0.6817	0.6679	0.6657
Subsample volatility level: Medium										
MSE	RW	0.3389	0.3483**	0.3247***	**0.3168*****	0.3173***	0.3217**	0.3238**	0.3268	0.3261
	EW	**0.3050**	0.3194***	0.3199***	0.3254***	0.3162**	0.3164*	0.3264***	0.3201*	0.3192
MAPE	RW	0.0435	0.0441**	0.0426***	**0.0422*****	0.0423***	0.0425***	0.0425**	0.0427	0.0427
	EW	**0.0413**	0.0422***	0.0423***	0.0426***	0.0422**	0.0421	0.0429***	0.0426*	0.0426*
Theil's U	RW	0.7320	0.7454	0.7143	**0.7069**	0.7085	0.7135	0.7163	0.7195	0.7181
	EW	**0.6955**	0.7174	0.7177	0.7229	0.7107	0.7051	0.7180	0.7088	0.7075
R^2	RW	-0.7972	-0.8468	-0.7218	**-0.6800**	-0.6828	-0.7059	-0.7173	-0.7331	-0.7293
	EW	**-2.2335**	-2.3852	-2.3908	-2.4491	-2.3519	-2.3543	-2.4594	-2.3932	-2.3840
Subsample volatility level: Low										
MSE	RW	0.6063	0.6180***	**0.5981***	0.6054	0.6113	0.6091	0.6131	0.6159	0.6156
	EW	0.6343	0.6647***	0.6658***	0.6512***	0.6621***	0.6289	0.6356	**0.6204**	0.6248
MAPE	RW	0.0534	0.0543***	**0.0534**	0.0538	0.0540	0.0538	0.0540	0.0542	0.0539
	EW	0.0550	0.0567***	0.0567***	0.0557***	0.0563***	0.0543*	0.0542*	0.0532***	**0.0532*****
Theil's U	RW	0.9231	0.9319	**0.9157**	0.9198	0.9233	0.9213	0.9238	0.9261	0.9258
	EW	0.9394	0.9620	0.9629	0.9515	0.9581	0.9329	0.9365	**0.9234**	0.9258
R^2	RW	0.6335	0.6264	**0.6384**	0.6340	0.6305	0.6318	0.6294	0.6277	0.6278

	EW	0.6339	0.6163	0.6157	0.6241	0.6178	0.6370	0.6332	**0.6419**	0.6393
Sample size		668	668	668	668	668	668	668	668	668

The table reports the forecasting performance of eight principal component analysis (PCA) models and the HAR model. Each PCA model includes the original HAR model independent variables (daily, weekly, monthly lagged realized volatility) and the top principal components extracted from daily word frequency data. PCA-1 indicates that the top 1 principal component has been included as an independent variable; PCA-2 indicates both top 2 principal components have been included as the independent variable; PCA-3 indicates all of the top 3 principal components have been included as independent variable etc.RW and EW stand for rolling windows and expanding windows, respectively; the rolling window is two years long. The significance of the forecasted value is indicated according to Diebold-Mariano statistics for the null of equal predictive accuracy of the word model and the HAR model under the absolute value (AV) loss function. ***, ** and * indicate the significant level of 1%, 5%, and 10% respectively. Updating frequency=22 indicates the parameters of models are re-estimated on a monthly basis. The table divides the entire sample period into three distinct subsamples based on the level of volatility: high, medium, and low volatility. This table examines the sensitivity of the results to changes in the volatility measure used for forecasting. Specifically, the main analysis used realized volatility (5-min) as the volatility measure, while the robustness check used realized kernel.

Table 18. Economic performance

Risk preference			1			
λ_0			$0.0273(1/4\mu)$	$0.0545(1/2\mu)$	$0.0818(3/4\mu)$	$0.1091(\mu)$
Market conditions			0.0063	0.0250	0.0563	0.1001
Panel A: Updating frequency=1						
		Forecasting skill	Utility gain			
log-HAR	RW	1.3199	0.0083	0.0330	0.0743	0.1321
	EW	1.1249	0.0070	0.0281	0.0633	0.1126
log-HAR-PCA-1	RW	1.4043	0.0088	0.0351	0.0790	0.1405
	EW	1.2196	0.0076	0.0305	0.0686	0.1220
log-HAR-PCA-2	RW	1.3524	0.0085	0.0338	0.0761	0.1353
	EW	1.2241	0.0077	0.0306	0.0689	0.1225
log-HAR-PCA-3	RW	1.3542	0.0085	0.0339	0.0762	0.1355
	EW	1.2485	0.0078	0.0312	0.0703	0.1249
log-HAR-PCA-4	RW	1.3736	0.0086	0.0344	0.0773	0.1374
	EW	1.2469	0.0078	0.0312	0.0702	0.1248
log-HAR-PCA-5	RW	1.3952	0.0087	0.0349	0.0785	0.1396
	EW	1.2736	0.0080	0.0319	0.0717	0.1274
log-HAR-PCA-6	RW	1.3971	0.0087	0.0350	0.0786	0.1398
	EW	1.3235	0.0083	0.0331	0.0745	0.1324
log-HAR-PCA-7	RW	1.4119	0.0088	0.0353	0.0795	0.1413
	EW	1.3030	0.0081	0.0326	0.0733	0.1304
log-HAR-PCA-8	RW	**1.4435**	**0.0090**	**0.0361**	**0.0812**	**0.1444**
	EW	1.3356	0.0084	0.0334	0.0752	0.1336
Panel B: Updating frequency=5						
		Forecasting skill	Advantage			
log-HAR	RW	1.3025	0.0081	0.0326	0.0733	0.1303
	EW	1.1173	0.0070	0.0280	0.0629	0.1118
log-HAR-PCA-1	RW	1.3881	0.0087	0.0347	0.0781	0.1389
	EW	1.2112	0.0076	0.0303	0.0682	0.1212
log-HAR-PCA-2	RW	1.3360	0.0084	0.0334	0.0752	0.1337
	EW	1.2148	0.0076	0.0304	0.0684	0.1216
log-HAR-PCA-3	RW	1.3315	0.0083	0.0333	0.0749	0.1332
	EW	1.2379	0.0077	0.0310	0.0697	0.1239
log-HAR-PCA-4	RW	1.3455	0.0084	0.0337	0.0757	0.1346
	EW	1.2328	0.0077	0.0308	0.0694	0.1234
log-HAR-PCA-5	RW	1.3670	0.0085	0.0342	0.0769	0.1368
	EW	1.2355	0.0077	0.0309	0.0695	0.1236
log-HAR-PCA-6	RW	1.3656	0.0085	0.0342	0.0769	0.1366
	EW	1.2786	0.0080	0.0320	0.0720	0.1279
log-HAR-PCA-7	RW	1.3624	0.0085	0.0341	0.0767	0.1363
	EW	1.2549	0.0078	0.0314	0.0706	0.1256
log-HAR-PCA-8	RW	**1.3838**	**0.0087**	**0.0346**	**0.0779**	**0.1385**
	EW	1.2779	0.0080	0.0320	0.0719	0.1279
Panel C: Updating frequency=22						
		Forecasting skill	Advantage			

log-HAR	RW	1.2640	0.0079	0.0316	0.0711	0.1265
	EW	1.1006	0.0069	0.0275	0.0619	0.1101
log-HAR-PCA-1	RW	**1.3429**	**0.0084**	**0.0336**	**0.0756**	**0.1344**
	EW	1.1906	0.0074	0.0298	0.0670	0.1191
log-HAR-PCA-2	RW	1.2832	0.0080	0.0321	0.0722	0.1284
	EW	1.1918	0.0075	0.0298	0.0671	0.1193
log-HAR-PCA-3	RW	1.2700	0.0079	0.0318	0.0715	0.1271
	EW	1.2160	0.0076	0.0304	0.0684	0.1217
log-HAR-PCA-4	RW	1.2746	0.0080	0.0319	0.0717	0.1275
	EW	1.2184	0.0076	0.0305	0.0686	0.1219
log-HAR-PCA-5	RW	1.2995	0.0081	0.0325	0.0731	0.1300
	EW	1.2122	0.0076	0.0303	0.0682	0.1213
log-HAR-PCA-6	RW	1.2864	0.0080	0.0322	0.0724	0.1287
	EW	1.2543	0.0078	0.0314	0.0706	0.1255
log-HAR-PCA-7	RW	1.2762	0.0080	0.0319	0.0718	0.1277
	EW	1.2249	0.0077	0.0306	0.0689	0.1226
log-HAR-PCA-8	RW	1.2790	0.0080	0.0320	0.0720	0.1280
	EW	1.2348	0.0077	0.0309	0.0695	0.1236

The table presents the economic significance of the conditional MV strategy that employs forecasting models to predict the realized variance of the S&P500 Index from 2014 to 2021. The utility gain metric assesses the economic significance, which incorporates three key elements: forecasting skill, risk preference, and market conditions. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 19. Leverages for the full sample period

θ			1			
λ_0			0.0273	0.0545	0.0818	0.1091
Market conditions			0.0063	0.0250	0.0563	0.1001
Panel A: Updating frequency=1						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	1.3199	2.4399	3.0452	3.6506	4.2559
	EW	1.1249	2.3504	2.8664	3.3823	3.8982
log-HAR-PCA-1	RW	1.4043	2.4786	3.1227	3.7667	4.4108
	EW	1.2196	2.3939	2.9532	3.5126	4.0719
log-HAR-PCA-2	RW	1.3524	2.4548	3.0751	3.6953	4.3156
	EW	1.2241	2.3959	2.9573	3.5187	4.0801
log-HAR-PCA-3	RW	1.3542	2.4556	3.0767	3.6978	4.3188
	EW	1.2485	2.4071	2.9797	3.5523	4.1249
log-HAR-PCA-4	RW	1.3736	2.4645	3.0945	3.7245	4.3544
	EW	1.2469	2.4064	2.9783	3.5501	4.1220
log-HAR-PCA-5	RW	1.3952	2.4744	3.1143	3.7542	4.3941
	EW	1.2736	2.4186	3.0027	3.5869	4.1710
log-HAR-PCA-6	RW	1.3971	2.4753	3.1161	3.7568	4.3976
	EW	1.3235	2.4416	3.0486	3.6556	4.2626
log-HAR-PCA-7	RW	1.4119	2.4821	3.1296	3.7772	4.4247
	EW	1.3030	2.4321	3.0298	3.6274	4.2250
log-HAR-PCA-8	RW	**1.4435**	**2.4966**	**3.1586**	**3.8206**	**4.4826**
	EW	1.3356	2.4471	3.0596	3.6721	4.2847
Panel B: Updating frequency=5						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	1.3025	2.4319	3.0292	3.6266	4.2239
	EW	1.1173	2.3470	2.8594	3.3719	3.8843
log-HAR-PCA-1	RW	1.3881	2.4712	3.1078	3.7445	4.3811
	EW	1.2112	2.3900	2.9455	3.5011	4.0566
log-HAR-PCA-2	RW	1.3360	2.4473	3.0600	3.6727	4.2854
	EW	1.2148	2.3917	2.9488	3.5059	4.0631
log-HAR-PCA-3	RW	1.3315	2.4452	3.0559	3.6666	4.2772
	EW	1.2379	2.4023	2.9700	3.5378	4.1055
log-HAR-PCA-4	RW	1.3455	2.4516	3.0687	3.6859	4.3030
	EW	1.2328	2.4000	2.9654	3.5308	4.0962
log-HAR-PCA-5	RW	1.3670	2.4615	3.0884	3.7154	4.3424
	EW	1.2355	2.4012	2.9678	3.5344	4.1010
log-HAR-PCA-6	RW	1.3656	2.4608	3.0872	3.7135	4.3398
	EW	1.2786	2.4210	3.0074	3.5938	4.1802
log-HAR-PCA-7	RW	1.3624	2.4594	3.0842	3.7090	4.3339
	EW	1.2549	2.4101	2.9856	3.5611	4.1367
log-HAR-PCA-8	RW	**1.3838**	**2.4692**	**3.1039**	**3.7386**	**4.3732**
	EW	1.2779	2.4206	3.0067	3.5928	4.1789
Panel C: Updating frequency=22						
		Forecasting skill	Leverage(mean)			

log-HAR	RW	1.2640	2.4143	2.9940	3.5737	4.1534
	EW	1.1006	2.3393	2.8441	3.3488	3.8536
log-HAR-PCA-1	RW	**1.3429**	**2.4504**	**3.0663**	**3.6822**	**4.2981**
	EW	1.1906	2.3806	2.9266	3.4727	4.0188
log-HAR-PCA-2	RW	1.2832	2.4231	3.0116	3.6001	4.1887
	EW	1.1918	2.3811	2.9277	3.4743	4.0208
log-HAR-PCA-3	RW	1.2700	2.4170	2.9995	3.5819	4.1644
	EW	1.2160	2.3922	2.9499	3.5076	4.0653
log-HAR-PCA-4	RW	1.2746	2.4191	3.0037	3.5883	4.1728
	EW	1.2184	2.3934	2.9522	3.5110	4.0698
log-HAR-PCA-5	RW	1.2995	2.4305	3.0265	3.6225	4.2185
	EW	1.2122	2.3905	2.9464	3.5024	4.0583
log-HAR-PCA-6	RW	1.2864	2.4245	3.0145	3.6045	4.1945
	EW	1.2543	2.4098	2.9851	3.5604	4.1356
log-HAR-PCA-7	RW	1.2762	2.4198	3.0051	3.5904	4.1757
	EW	1.2249	2.3963	2.9581	3.5199	4.0817
log-HAR-PCA-8	RW	1.2790	2.4211	3.0077	3.5943	4.1809
	EW	1.2348	2.4009	2.9672	3.5335	4.0999

The table presents the mean leverage required in achieving utility gains of Table 16 by implementing forecasting models to predict the realized variance of the S&P500 Index from 2014 to 2021. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 20. Sub-sample economic performance-High volatility regime

Risk preference			0.0122			
λ_0			-0.1100$(1/4\mu)$	-0.2200$(1/2\mu)$	-0.3300$(3/4\mu)$	-0.4400(μ)
Market condition			0.0410	0.1639	0.3688	0.6557
Panel A: Updating frequency=1						
		Forecasting skill	Utility gain			
log-HAR	RW	1.0894	0.0005	0.0022	0.0049	0.0087
	EW	0.9322	0.0005	0.0019	0.0042	0.0075
log-HAR-PCA-1	RW	1.1746	0.0006	0.0023	0.0053	0.0094
	EW	1.0255	0.0005	0.0020	0.0046	0.0082
log-HAR-PCA-2	RW	1.1379	0.0006	0.0023	0.0051	0.0091
	EW	1.0252	0.0005	0.0020	0.0046	0.0082
log-HAR-PCA-3	RW	1.1897	0.0006	0.0024	0.0053	0.0095
	EW	1.0494	0.0005	0.0021	0.0047	0.0084
log-HAR-PCA-4	RW	1.2132	0.0006	0.0024	0.0055	0.0097
	EW	1.0863	0.0005	0.0022	0.0049	0.0087
log-HAR-PCA-5	RW	1.2431	0.0006	0.0025	0.0056	0.0099
	EW	1.1365	0.0006	0.0023	0.0051	0.0091
log-HAR-PCA-6	RW	1.2408	0.0006	0.0025	0.0056	0.0099
	EW	1.1797	0.0006	0.0024	0.0053	0.0094
log-HAR-PCA-7	RW	1.2425	0.0006	0.0025	0.0056	0.0099
	EW	1.1772	0.0006	0.0024	0.0053	0.0094
log-HAR-PCA-8	RW	**1.2768**	**0.0006**	**0.0026**	**0.0057**	**0.0102**
	EW	1.2266	0.0006	0.0025	0.0055	0.0098
Panel B: Updating frequency=5						
		Forecasting skill	Advantage			
log-HAR	RW	1.0705	0.0005	0.0021	0.0048	0.0086
	EW	0.9250	0.0005	0.0018	0.0042	0.0074
log-HAR-PCA-1	RW	1.1577	0.0006	0.0023	0.0052	0.0093
	EW	1.0175	0.0005	0.0020	0.0046	0.0081
log-HAR-PCA-2	RW	1.1190	0.0006	0.0022	0.0050	0.0089
	EW	1.0185	0.0005	0.0020	0.0046	0.0081
log-HAR-PCA-3	RW	1.1643	0.0006	0.0023	0.0052	0.0093
	EW	1.0417	0.0005	0.0021	0.0047	0.0083
log-HAR-PCA-4	RW	1.1791	0.0006	0.0024	0.0053	0.0094
	EW	1.0766	0.0005	0.0022	0.0048	0.0086
log-HAR-PCA-5	RW	1.2045	0.0006	0.0024	0.0054	0.0096
	EW	1.0785	0.0005	0.0022	0.0048	0.0086
log-HAR-PCA-6	RW	1.1955	0.0006	0.0024	0.0054	0.0096
	EW	1.1213	0.0006	0.0022	0.0050	0.0090
log-HAR-PCA-7	RW	1.1841	0.0006	0.0024	0.0053	0.0095
	EW	1.1169	0.0006	0.0022	0.0050	0.0089
log-HAR-PCA-8	RW	**1.2086**	**0.0006**	**0.0024**	**0.0054**	**0.0097**
	EW	1.1542	0.0006	0.0023	0.0052	0.0092
Panel C: Updating frequency=22						
		Forecasting skill	Advantage			

log-HAR	RW	1.0206	0.0005	0.0020	0.0046	0.0082
	EW	0.9065	0.0005	0.0018	0.0041	0.0072
log-HAR-PCA-1	RW	1.0837	0.0005	0.0022	0.0049	0.0087
	EW	0.9919	0.0005	0.0020	0.0045	0.0079
log-HAR-PCA-2	RW	1.0459	0.0005	0.0021	0.0047	0.0084
	EW	0.9942	0.0005	0.0020	0.0045	0.0079
log-HAR-PCA-3	RW	1.0772	0.0005	0.0022	0.0048	0.0086
	EW	1.0154	0.0005	0.0020	0.0046	0.0081
log-HAR-PCA-4	RW	1.0854	0.0005	0.0022	0.0049	0.0087
	EW	1.0421	0.0005	0.0021	0.0047	0.0083
log-HAR-PCA-5	RW	**1.1086**	**0.0006**	**0.0022**	**0.0050**	**0.0089**
	EW	1.0210	0.0005	0.0020	0.0046	0.0082
log-HAR-PCA-6	RW	1.0822	0.0005	0.0022	0.0049	0.0087
	EW	1.0605	0.0005	0.0021	0.0048	0.0085
log-HAR-PCA-7	RW	1.0470	0.0005	0.0021	0.0047	0.0084
	EW	1.0483	0.0005	0.0021	0.0047	0.0084
log-HAR-PCA-8	RW	1.0484	0.0005	0.0021	0.0047	0.0084
	EW	1.0674	0.0005	0.0021	0.0048	0.0085

The table presents the economic significance of the conditional MV strategy that employs forecasting models to predict the realized variance of high volatility subsamples of the S&P500 Index from 2014 to 2021. The utility gain metric assesses the economic significance, which incorporates three key elements: forecasting skill, risk preference, and market conditions. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 21. Sub-sample economic performance-Medium volatility regime

Risk preference			0.0122			
λ_0			$0.0866(1/4\mu)$	$0.1731(1/2\mu)$	$0.2597(3/4\mu)$	$0.3463(\mu)$
Market condition			0.1822	0.7286	1.6394	2.9145
Panel A: Updating frequency=1						
		Forecasting skill	Advantage			
log-HAR	RW	0.3337	0.0007	0.0030	0.0067	0.0119
	EW	0.2927	0.0007	0.0026	0.0059	0.0104
log-HAR-PCA-1	RW	**0.3437**	**0.0008**	**0.0031**	**0.0069**	**0.0122**
	EW	0.3100	0.0007	0.0028	0.0062	0.0110
log-HAR-PCA-2	RW	0.3286	0.0007	0.0029	0.0066	0.0117
	EW	0.3120	0.0007	0.0028	0.0062	0.0111
log-HAR-PCA-3	RW	0.3251	0.0007	0.0029	0.0065	0.0116
	EW	0.3169	0.0007	0.0028	0.0063	0.0113
log-HAR-PCA-4	RW	0.3278	0.0007	0.0029	0.0066	0.0116
	EW	0.3034	0.0007	0.0027	0.0061	0.0108
log-HAR-PCA-5	RW	0.3268	0.0007	0.0029	0.0065	0.0116
	EW	0.3107	0.0007	0.0028	0.0062	0.0110
log-HAR-PCA-6	RW	0.3296	0.0007	0.0029	0.0066	0.0117
	EW	0.3207	0.0007	0.0028	0.0064	0.0114
log-HAR-PCA-7	RW	0.3349	0.0007	0.0030	0.0067	0.0119
	EW	0.3228	0.0007	0.0029	0.0065	0.0115
log-HAR-PCA-8	RW	0.3365	0.0007	0.0030	0.0067	0.0120
	EW	0.3242	0.0007	0.0029	0.0065	0.0115
Panel B: Updating frequency=5						
		Forecasting skill	Advantage			
log-HAR	RW	0.3323	0.0007	0.0030	0.0066	0.0118
	EW	0.2919	0.0006	0.0026	0.0058	0.0104
log-HAR-PCA-1	RW	0.3424	0.0008	0.0030	0.0068	0.0122
	EW	0.3092	0.0007	0.0027	0.0062	0.0110
log-HAR-PCA-2	RW	0.3271	0.0007	0.0029	0.0065	0.0116
	EW	0.3110	0.0007	0.0028	0.0062	0.0111
log-HAR-PCA-3	RW	0.3254	0.0007	0.0029	0.0065	0.0116
	EW	0.3187	0.0007	0.0028	0.0064	0.0113
log-HAR-PCA-4	RW	0.3288	0.0007	0.0029	0.0066	0.0117
	EW	0.3030	0.0007	0.0027	0.0061	0.0108
log-HAR-PCA-5	RW	0.3274	0.0007	0.0029	0.0065	0.0116
	EW	0.3097	0.0007	0.0028	0.0062	0.0110
log-HAR-PCA-6	RW	0.3378	0.0008	0.0030	0.0068	0.0120
	EW	0.3227	0.0007	0.0029	0.0064	0.0115
log-HAR-PCA-7	RW	0.3406	0.0008	0.0030	0.0068	0.0121
	EW	0.3241	0.0007	0.0029	0.0065	0.0115
log-HAR-PCA-8	RW	**0.3434**	**0.0008**	**0.0031**	**0.0069**	**0.0122**
	EW	0.3242	0.0007	0.0029	0.0065	0.0115
Panel C: Updating frequency=22						
		Forecasting skill	Advantage			

log-HAR	RW	0.3293	0.0007	0.0029	0.0066	0.0117
	EW	0.2899	0.0006	0.0026	0.0058	0.0103
log-HAR-PCA-1	RW	**0.3416**	**0.0008**	**0.0030**	**0.0068**	**0.0121**
	EW	0.3075	0.0007	0.0027	0.0061	0.0109
log-HAR-PCA-2	RW	0.3229	0.0007	0.0029	0.0065	0.0115
	EW	0.3077	0.0007	0.0027	0.0062	0.0109
log-HAR-PCA-3	RW	0.3219	0.0007	0.0029	0.0064	0.0114
	EW	0.3145	0.0007	0.0028	0.0063	0.0112
log-HAR-PCA-4	RW	0.3222	0.0007	0.0029	0.0064	0.0114
	EW	0.3031	0.0007	0.0027	0.0061	0.0108
log-HAR-PCA-5	RW	0.3228	0.0007	0.0029	0.0065	0.0115
	EW	0.3106	0.0007	0.0028	0.0062	0.0110
log-HAR-PCA-6	RW	0.3293	0.0007	0.0029	0.0066	0.0117
	EW	0.3244	0.0007	0.0029	0.0065	0.0115
log-HAR-PCA-7	RW	0.3360	0.0007	0.0030	0.0067	0.0119
	EW	0.3261	0.0007	0.0029	0.0065	0.0116
log-HAR-PCA-8	RW	0.3373	0.0007	0.0030	0.0067	0.0120
	EW	0.3258	0.0007	0.0029	0.0065	0.0116

The table presents the economic significance of the conditional MV strategy that employs forecasting models to predict the realized variance of medium volatility subsamples of the S&P500 Index from 2014 to 2021. The utility gain metric assesses the economic significance, which incorporates three key elements: forecasting skill, risk preference, and market conditions. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 22. Sub-sample economic performance-Low volatility regime

Risk preference			0.0122			
λ_0			$0.1059(1/4\mu)$	$0.2118(1/2\mu)$	$0.3178(3/4\mu)$	$0.4237(\mu)$
Market condition			0.8365	3.3459	7.5282	13.3835
Panel A: Updating frequency=1						
		Forecasting skill	Advantage			
log-HAR	RW	0.3236	0.0033	0.0132	0.0297	0.0528
	EW	0.2788	0.0028	0.0114	0.0256	0.0455
log-HAR-PCA-1	RW	**0.3342**	**0.0034**	**0.0136**	**0.0307**	**0.0545**
	EW	0.2890	0.0029	0.0118	0.0265	0.0472
log-HAR-PCA-2	RW	0.3218	0.0033	0.0131	0.0295	0.0525
	EW	0.2904	0.0030	0.0118	0.0267	0.0474
log-HAR-PCA-3	RW	0.3018	0.0031	0.0123	0.0277	0.0492
	EW	0.2860	0.0029	0.0117	0.0263	0.0467
log-HAR-PCA-4	RW	0.3006	0.0031	0.0123	0.0276	0.0490
	EW	0.2790	0.0028	0.0114	0.0256	0.0455
log-HAR-PCA-5	RW	0.3001	0.0031	0.0122	0.0275	0.0490
	EW	0.2836	0.0029	0.0116	0.0260	0.0463
log-HAR-PCA-6	RW	0.3022	0.0031	0.0123	0.0277	0.0493
	EW	0.2835	0.0029	0.0116	0.0260	0.0463
log-HAR-PCA-7	RW	0.3066	0.0031	0.0125	0.0281	0.0500
	EW	0.2717	0.0028	0.0111	0.0249	0.0443
log-HAR-PCA-8	RW	0.3065	0.0031	0.0125	0.0281	0.0500
	EW	0.2722	0.0028	0.0111	0.0250	0.0444
Panel B: Updating frequency=5						
		Forecasting skill	Advantage			
log-HAR	RW	0.3215	0.0033	0.0131	0.0295	0.0525
	EW	0.2776	0.0028	0.0113	0.0255	0.0453
log-HAR-PCA-1	RW	**0.3320**	**0.0034**	**0.0135**	**0.0305**	**0.0542**
	EW	0.2878	0.0029	0.0117	0.0264	0.0470
log-HAR-PCA-2	RW	0.3204	0.0033	0.0131	0.0294	0.0523
	EW	0.2891	0.0029	0.0118	0.0265	0.0472
log-HAR-PCA-3	RW	0.2999	0.0031	0.0122	0.0275	0.0489
	EW	0.2856	0.0029	0.0117	0.0262	0.0466
log-HAR-PCA-4	RW	0.2978	0.0030	0.0121	0.0273	0.0486
	EW	0.2798	0.0029	0.0114	0.0257	0.0457
log-HAR-PCA-5	RW	0.2975	0.0030	0.0121	0.0273	0.0486
	EW	0.2851	0.0029	0.0116	0.0262	0.0465
log-HAR-PCA-6	RW	0.3015	0.0031	0.0123	0.0277	0.0492
	EW	0.2837	0.0029	0.0116	0.0260	0.0463
log-HAR-PCA-7	RW	0.3056	0.0031	0.0125	0.0281	0.0499
	EW	0.2727	0.0028	0.0111	0.0250	0.0445
log-HAR-PCA-8	RW	0.3049	0.0031	0.0124	0.0280	0.0497
	EW	0.2726	0.0028	0.0111	0.0250	0.0445
Panel C: Updating frequency=22						
		Forecasting skill	Advantage			

log-HAR	RW	0.3183	0.0032	0.0130	0.0292	0.0519
	EW	0.2756	0.0028	0.0112	0.0253	0.0450
log-HAR-PCA-1	RW	**0.3303**	**0.0034**	**0.0135**	**0.0303**	**0.0539**
	EW	0.2860	0.0029	0.0117	0.0263	0.0467
log-HAR-PCA-2	RW	0.3183	0.0032	0.0130	0.0292	0.0519
	EW	0.2875	0.0029	0.0117	0.0264	0.0469
log-HAR-PCA-3	RW	0.2988	0.0030	0.0122	0.0274	0.0488
	EW	0.3015	0.0031	0.0123	0.0277	0.0492
log-HAR-PCA-4	RW	0.2951	0.0030	0.0120	0.0271	0.0481
	EW	0.2996	0.0031	0.0122	0.0275	0.0489
log-HAR-PCA-5	RW	0.2977	0.0030	0.0121	0.0273	0.0486
	EW	0.3066	0.0031	0.0125	0.0281	0.0500
log-HAR-PCA-6	RW	0.3044	0.0031	0.0124	0.0279	0.0497
	EW	0.3135	0.0032	0.0128	0.0288	0.0512
log-HAR-PCA-7	RW	0.3072	0.0031	0.0125	0.0282	0.0501
	EW	0.3025	0.0031	0.0123	0.0278	0.0494
log-HAR-PCA-8	RW	0.3067	0.0031	0.0125	0.0281	0.0500
	EW	0.3064	0.0031	0.0125	0.0281	0.0500

The table presents the economic significance of the conditional MV strategy that employs forecasting models to predict the realized variance of low volatility subsamples of the S&P500 Index from 2014 to 2021. The utility gain metric assesses the economic significance, which incorporates three key elements: forecasting skill, risk preference, and market conditions. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 23. Leverages for the sub-sample period-High volatility regime

θ			82.0175			
λ_0			-0.1100	-0.2200	-0.3300	-0.4400
Market conditions			0.0410	0.1639	0.3688	0.6557
Panel A: Updating frequency=1						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	1.0894	-0.0462	-0.0561	-0.0660	-0.0759
	EW	0.9322	-0.0448	-0.0533	-0.0617	-0.0702
log-HAR-PCA-1	RW	1.1746	-0.0470	-0.0577	-0.0684	-0.0790
	EW	1.0255	-0.0457	-0.0550	-0.0643	-0.0736
log-HAR-PCA-2	RW	1.1379	-0.0467	-0.0570	-0.0674	-0.0777
	EW	1.0252	-0.0457	-0.0550	-0.0643	-0.0736
log-HAR-PCA-3	RW	1.1897	-0.0471	-0.0580	-0.0688	-0.0796
	EW	1.0494	-0.0459	-0.0554	-0.0649	-0.0745
log-HAR-PCA-4	RW	1.2132	-0.0474	-0.0584	-0.0694	-0.0804
	EW	1.0863	-0.0462	-0.0561	-0.0659	-0.0758
log-HAR-PCA-5	RW	1.2431	-0.0476	-0.0589	-0.0702	-0.0815
	EW	1.1365	-0.0467	-0.0570	-0.0673	-0.0776
log-HAR-PCA-6	RW	1.2408	-0.0476	-0.0589	-0.0702	-0.0814
	EW	1.1797	-0.0471	-0.0578	-0.0685	-0.0792
log-HAR-PCA-7	RW	1.2425	-0.0476	-0.0589	-0.0702	-0.0815
	EW	1.1772	-0.0470	-0.0577	-0.0684	-0.0791
log-HAR-PCA-8	RW	**1.2768**	**-0.0479**	**-0.0595**	**-0.0711**	**-0.0827**
	EW	1.2266	-0.0475	-0.0586	-0.0698	-0.0809
Panel B: Updating frequency=5						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	1.0705	-0.0461	-0.0558	-0.0655	-0.0752
	EW	0.9250	-0.0447	-0.0531	-0.0616	-0.0700
log-HAR-PCA-1	RW	1.1577	-0.0469	-0.0574	-0.0679	-0.0784
	EW	1.0175	-0.0456	-0.0548	-0.0641	-0.0733
log-HAR-PCA-2	RW	1.1190	-0.0465	-0.0567	-0.0668	-0.0770
	EW	1.0185	-0.0456	-0.0548	-0.0641	-0.0734
log-HAR-PCA-3	RW	1.1643	-0.0469	-0.0575	-0.0681	-0.0786
	EW	1.0417	-0.0458	-0.0553	-0.0647	-0.0742
log-HAR-PCA-4	RW	1.1791	-0.0471	-0.0578	-0.0685	-0.0792
	EW	1.0766	-0.0461	-0.0559	-0.0657	-0.0755
log-HAR-PCA-5	RW	1.2045	-0.0473	-0.0582	-0.0692	-0.0801
	EW	1.0785	-0.0461	-0.0559	-0.0657	-0.0755
log-HAR-PCA-6	RW	1.1955	-0.0472	-0.0581	-0.0689	-0.0798
	EW	1.1213	-0.0465	-0.0567	-0.0669	-0.0771
log-HAR-PCA-7	RW	1.1841	-0.0471	-0.0579	-0.0686	-0.0794
	EW	1.1169	-0.0465	-0.0566	-0.0668	-0.0769
log-HAR-PCA-8	RW	**1.2086**	**-0.0473**	**-0.0583**	**-0.0693**	**-0.0803**
	EW	1.1542	-0.0468	-0.0573	-0.0678	-0.0783
Panel C: Updating frequency=22						
		Forecasting skill	Leverage(mean)			

log-HAR	RW	1.0206	-0.0456	-0.0549	-0.0642	-0.0734
	EW	0.9065	-0.0446	-0.0528	-0.0610	-0.0693
log-HAR-PCA-1	RW	1.0837	-0.0462	-0.0560	-0.0659	-0.0757
	EW	0.9919	-0.0454	-0.0544	-0.0634	-0.0724
log-HAR-PCA-2	RW	1.0459	-0.0458	-0.0553	-0.0648	-0.0743
	EW	0.9942	-0.0454	-0.0544	-0.0634	-0.0725
log-HAR-PCA-3	RW	1.0772	-0.0461	-0.0559	-0.0657	-0.0755
	EW	1.0154	-0.0456	-0.0548	-0.0640	-0.0732
log-HAR-PCA-4	RW	1.0854	-0.0462	-0.0561	-0.0659	-0.0758
	EW	1.0421	-0.0458	-0.0553	-0.0647	-0.0742
log-HAR-PCA-5	RW	**1.1086**	**-0.0464**	**-0.0565**	**-0.0666**	**-0.0766**
	EW	1.0210	-0.0456	-0.0549	-0.0642	-0.0734
log-HAR-PCA-6	RW	1.0822	-0.0462	-0.0560	-0.0658	-0.0757
	EW	1.0605	-0.0460	-0.0556	-0.0652	-0.0749
log-HAR-PCA-7	RW	1.0470	-0.0459	-0.0554	-0.0649	-0.0744
	EW	1.0483	-0.0459	-0.0554	-0.0649	-0.0744
log-HAR-PCA-8	RW	1.0484	-0.0459	-0.0554	-0.0649	-0.0744
	EW	1.0674	-0.0460	-0.0557	-0.0654	-0.0751

The table presents the mean leverage required in achieving utility gains of Table 18 by implementing forecasting models to predict the realized variance of the high volatility subsample of the S&P500 Index from 2014 to 2021. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 24. Leverages for the sub-sample period-Medium volatility regime

θ			82.0175			
λ_0			0.0866	0.1731	0.2597	0.3463
Market conditions			0.1822	0.7286	1.6394	2.9145
Panel A: Updating frequency=1						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	0.3337	0.2224	0.2395	0.2566	0.2737
	EW	0.2927	0.2203	0.2353	0.2503	0.2653
log-HAR-PCA-1	RW	**0.3437**	**0.2229**	**0.2405**	**0.2581**	**0.2758**
	EW	0.3100	0.2211	0.2370	0.2530	0.2689
log-HAR-PCA-2	RW	0.3286	0.2221	0.2390	0.2558	0.2727
	EW	0.3120	0.2212	0.2373	0.2533	0.2693
log-HAR-PCA-3	RW	0.3251	0.2219	0.2386	0.2553	0.2720
	EW	0.3169	0.2215	0.2378	0.2540	0.2703
log-HAR-PCA-4	RW	0.3278	0.2221	0.2389	0.2557	0.2725
	EW	0.3034	0.2208	0.2364	0.2519	0.2675
log-HAR-PCA-5	RW	0.3268	0.2220	0.2388	0.2555	0.2723
	EW	0.3107	0.2212	0.2371	0.2531	0.2690
log-HAR-PCA-6	RW	0.3296	0.2221	0.2391	0.2560	0.2729
	EW	0.3207	0.2217	0.2381	0.2546	0.2711
log-HAR-PCA-7	RW	0.3349	0.2224	0.2396	0.2568	0.2740
	EW	0.3228	0.2218	0.2384	0.2549	0.2715
log-HAR-PCA-8	RW	0.3365	0.2225	0.2398	0.2570	0.2743
	EW	0.3242	0.2219	0.2385	0.2551	0.2718
Panel B: Updating frequency=5						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	0.3323	0.2223	0.2393	0.2564	0.2734
	EW	0.2919	0.2202	0.2352	0.2502	0.2652
log-HAR-PCA-1	RW	0.3424	0.2228	0.2404	0.2579	0.2755
	EW	0.3092	0.2211	0.2370	0.2528	0.2687
log-HAR-PCA-2	RW	0.3271	0.2220	0.2388	0.2556	0.2724
	EW	0.3110	0.2212	0.2372	0.2531	0.2691
log-HAR-PCA-3	RW	0.3254	0.2219	0.2386	0.2553	0.2720
	EW	0.3187	0.2216	0.2379	0.2543	0.2706
log-HAR-PCA-4	RW	0.3288	0.2221	0.2390	0.2558	0.2727
	EW	0.3030	0.2208	0.2363	0.2519	0.2674
log-HAR-PCA-5	RW	0.3274	0.2220	0.2388	0.2556	0.2724
	EW	0.3097	0.2211	0.2370	0.2529	0.2688
log-HAR-PCA-6	RW	0.3378	0.2226	0.2399	0.2572	0.2746
	EW	0.3227	0.2218	0.2384	0.2549	0.2715
log-HAR-PCA-7	RW	0.3406	0.2227	0.2402	0.2577	0.2751
	EW	0.3241	0.2219	0.2385	0.2551	0.2718
log-HAR-PCA-8	RW	**0.3434**	**0.2229**	**0.2405**	**0.2581**	**0.2757**
	EW	0.3242	0.2219	0.2385	0.2551	0.2718
Panel C: Updating frequency=22						
		Forecasting skill	Leverage(mean)			

log-HAR	RW	0.3293	0.2221	0.2390	0.2559	0.2728
	EW	0.2899	0.2201	0.2350	0.2499	0.2647
log-HAR-PCA-1	RW	**0.3416**	**0.2228**	**0.2403**	**0.2578**	**0.2754**
	EW	0.3075	0.2210	0.2368	0.2526	0.2683
log-HAR-PCA-2	RW	0.3229	0.2218	0.2384	0.2549	0.2715
	EW	0.3077	0.2210	0.2368	0.2526	0.2684
log-HAR-PCA-3	RW	0.3219	0.2218	0.2383	0.2548	0.2713
	EW	0.3145	0.2214	0.2375	0.2536	0.2698
log-HAR-PCA-4	RW	0.3222	0.2218	0.2383	0.2548	0.2714
	EW	0.3031	0.2208	0.2363	0.2519	0.2674
log-HAR-PCA-5	RW	0.3228	0.2218	0.2384	0.2549	0.2715
	EW	0.3106	0.2212	0.2371	0.2531	0.2690
log-HAR-PCA-6	RW	0.3293	0.2221	0.2390	0.2559	0.2728
	EW	0.3244	0.2219	0.2385	0.2552	0.2718
log-HAR-PCA-7	RW	0.3360	0.2225	0.2397	0.2570	0.2742
	EW	0.3261	0.2220	0.2387	0.2554	0.2722
log-HAR-PCA-8	RW	0.3373	0.2225	0.2398	0.2572	0.2745
	EW	0.3258	0.2220	0.2387	0.2554	0.2721

The table presents the mean leverage required in achieving utility gains of Table 19 by implementing forecasting models to predict the realized variance of the medium volatility subsample of the S&P500 Index from 2014 to 2021. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

Table 25. Leverages for the sub-sample period-Low volatility regime

θ			82.0175			
λ_0			0.1059	0.2118	0.3178	0.4237
Market conditions			0.8365	3.3459	7.5282	13.3835
Panel A: Updating frequency=1						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	0.3236	0.8326	0.8949	0.9572	1.0196
	EW	0.2788	0.8240	0.8777	0.9314	0.9851
log-HAR-PCA-1	RW	**0.3342**	**0.8347**	**0.8990**	**0.9634**	**1.0277**
	EW	0.2890	0.8259	0.8816	0.9372	0.9929
log-HAR-PCA-2	RW	0.3218	0.8323	0.8943	0.9562	1.0182
	EW	0.2904	0.8262	0.8821	0.9381	0.9940
log-HAR-PCA-3	RW	0.3018	0.8284	0.8865	0.9446	1.0028
	EW	0.2860	0.8254	0.8805	0.9355	0.9906
log-HAR-PCA-4	RW	0.3006	0.8282	0.8861	0.9439	1.0018
	EW	0.2790	0.8240	0.8778	0.9315	0.9852
log-HAR-PCA-5	RW	0.3001	0.8281	0.8859	0.9437	1.0014
	EW	0.2836	0.8249	0.8795	0.9341	0.9888
log-HAR-PCA-6	RW	0.3022	0.8285	0.8867	0.9449	1.0031
	EW	0.2835	0.8249	0.8795	0.9341	0.9887
log-HAR-PCA-7	RW	0.3066	0.8293	0.8884	0.9474	1.0065
	EW	0.2717	0.8226	0.8749	0.9272	0.9796
log-HAR-PCA-8	RW	0.3065	0.8293	0.8884	0.9474	1.0064
	EW	0.2722	0.8227	0.8751	0.9276	0.9800
Panel B: Updating frequency=5						
		Forecasting skill	Leverage(mean)			
log-HAR	RW	0.3215	0.8322	0.8941	0.9560	1.0179
	EW	0.2776	0.8238	0.8772	0.9307	0.9842
log-HAR-PCA-1	RW	**0.3320**	**0.8342**	**0.8982**	**0.9621**	**1.0261**
	EW	0.2878	0.8257	0.8812	0.9366	0.9920
log-HAR-PCA-2	RW	0.3204	0.8320	0.8937	0.9554	1.0171
	EW	0.2891	0.8260	0.8816	0.9373	0.9930
log-HAR-PCA-3	RW	0.2999	0.8280	0.8858	0.9435	1.0013
	EW	0.2856	0.8253	0.8803	0.9353	0.9903
log-HAR-PCA-4	RW	0.2978	0.8276	0.8850	0.9423	0.9997
	EW	0.2798	0.8242	0.8781	0.9320	0.9858
log-HAR-PCA-5	RW	0.2975	0.8276	0.8849	0.9422	0.9995
	EW	0.2851	0.8252	0.8801	0.9350	0.9899
log-HAR-PCA-6	RW	0.3015	0.8284	0.8864	0.9445	1.0025
	EW	0.2837	0.8249	0.8796	0.9342	0.9888
log-HAR-PCA-7	RW	0.3056	0.8292	0.8880	0.9469	1.0057
	EW	0.2727	0.8228	0.8753	0.9278	0.9803
log-HAR-PCA-8	RW	0.3049	0.8290	0.8877	0.9464	1.0051
	EW	0.2726	0.8228	0.8753	0.9278	0.9803
Panel C: Updating frequency=22						
		Forecasting skill	Leverage(mean)			

log-HAR	RW	0.3183	0.8316	0.8929	0.9542	1.0155
	EW	0.2756	0.8234	0.8764	0.9295	0.9826
log-HAR-PCA-1	RW	**0.3303**	**0.8339**	**0.8975**	**0.9611**	**1.0248**
	EW	0.2860	0.8254	0.8805	0.9356	0.9906
log-HAR-PCA-2	RW	0.3183	0.8316	0.8929	0.9542	1.0155
	EW	0.2875	0.8257	0.8810	0.9364	0.9918
log-HAR-PCA-3	RW	0.2988	0.8278	0.8854	0.9429	1.0005
	EW	0.3015	0.8284	0.8864	0.9445	1.0025
log-HAR-PCA-4	RW	0.2951	0.8271	0.8839	0.9408	0.9976
	EW	0.2996	0.8280	0.8857	0.9434	1.0011
log-HAR-PCA-5	RW	0.2977	0.8276	0.8850	0.9423	0.9996
	EW	0.3066	0.8293	0.8884	0.9474	1.0065
log-HAR-PCA-6	RW	0.3044	0.8289	0.8875	0.9461	1.0048
	EW	0.3135	0.8307	0.8911	0.9514	1.0118
log-HAR-PCA-7	RW	0.3072	0.8295	0.8886	0.9478	1.0069
	EW	0.3025	0.8285	0.8868	0.9450	1.0033
log-HAR-PCA-8	RW	0.3067	0.8294	0.8884	0.9475	1.0065
	EW	0.3064	0.8293	0.8883	0.9473	1.0063

The table presents the mean leverage required in achieving utility gains of Table 20 by implementing forecasting models to predict the realized variance of the low volatility subsample of the S&P500 Index from 2014 to 2021. Forecasting skill is calculated as $\exp[C^2] - 1$, while risk preference $(1/\theta)$ is assumed to be 1. Market conditions $(\lambda_0^2/4\sigma^2)$ are determined by a parameter λ_0 that ranges from $1/4\mu$ to μ, where μ represents the excess return of the S&P500 Index over the 10-year Treasury bond yield and σ^2 is the variance of such excess returns. RW and EW stand for rolling windows and expanding windows, respectively, and the rolling window is two years long.

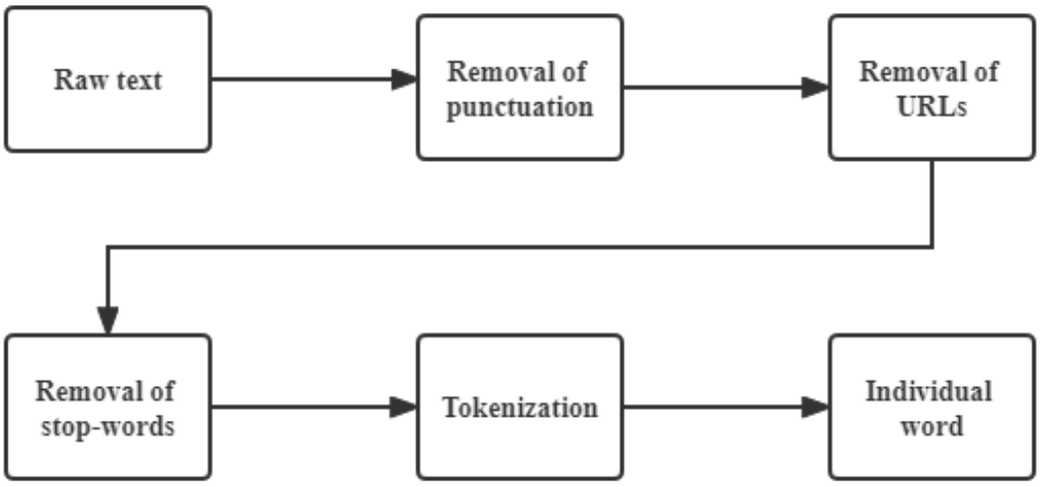

Figure 1: Preprocess steps of Twitter texts

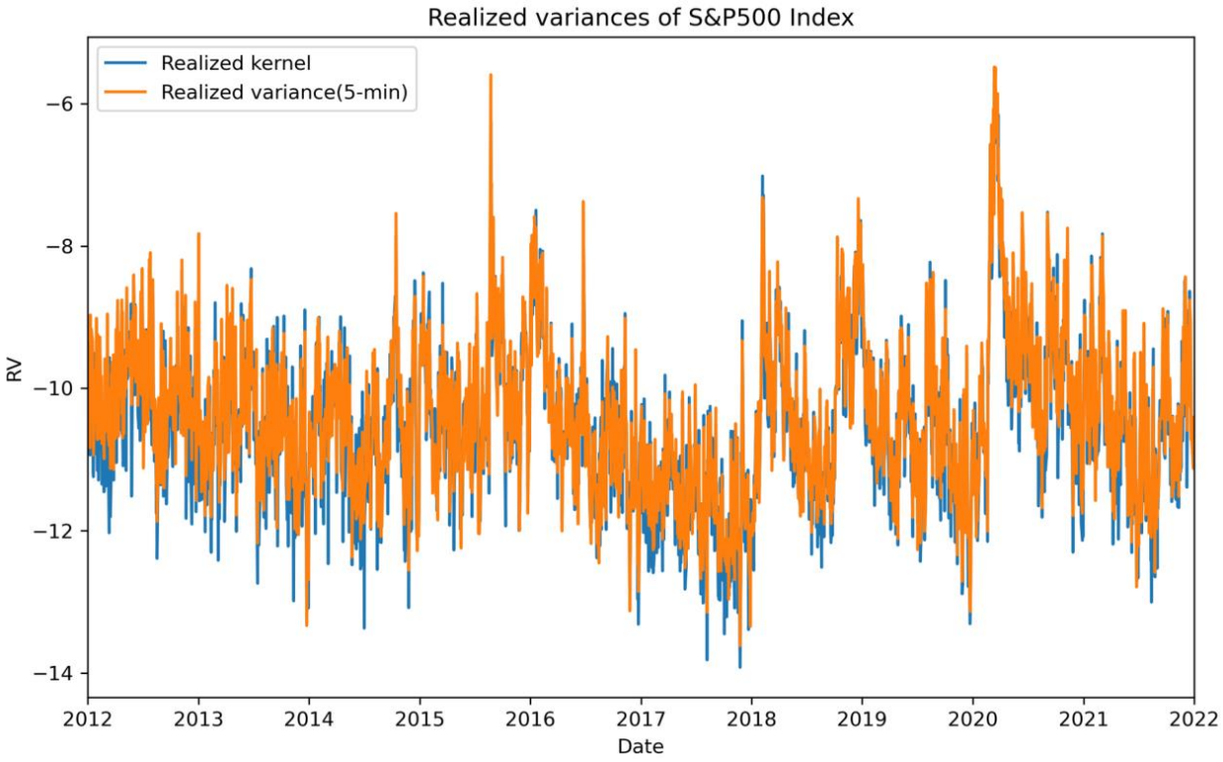

Figure 2. S&P500 Index realized variances series (log).

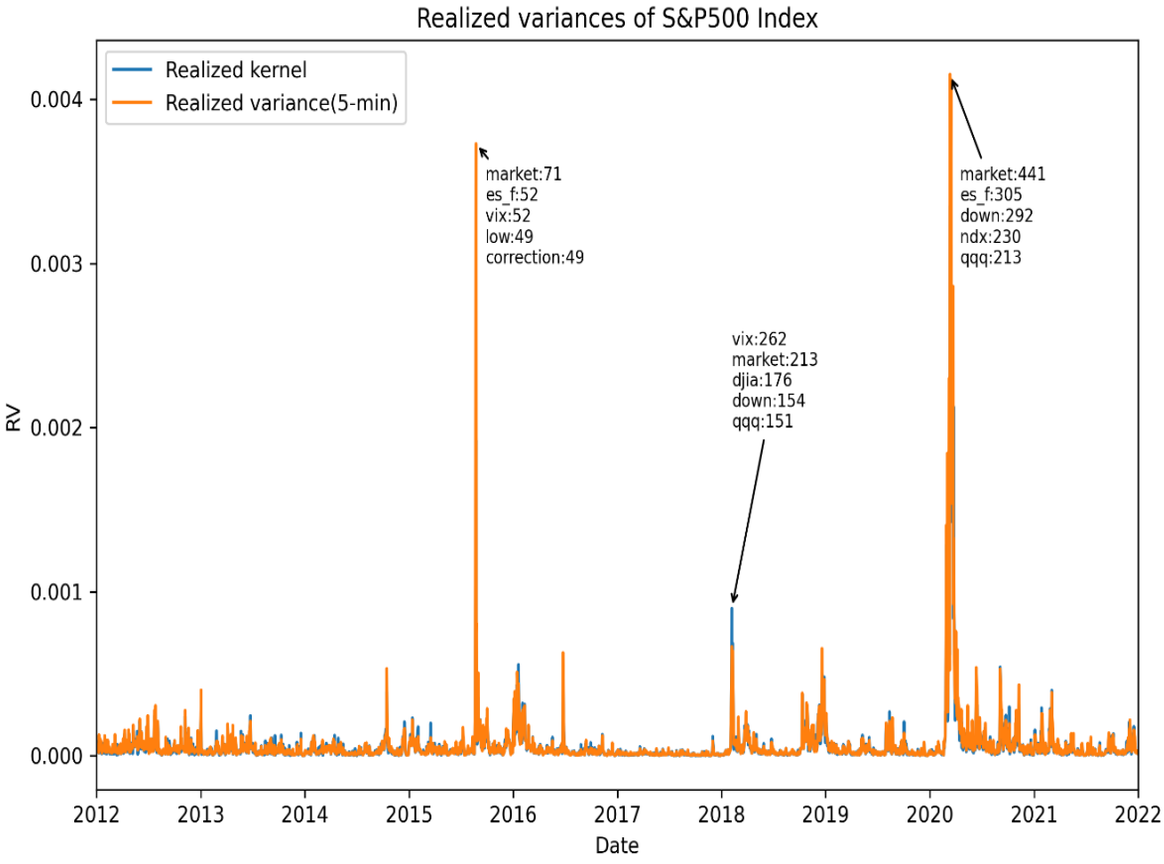

Figure 3. S&P500 Index realized variances series.

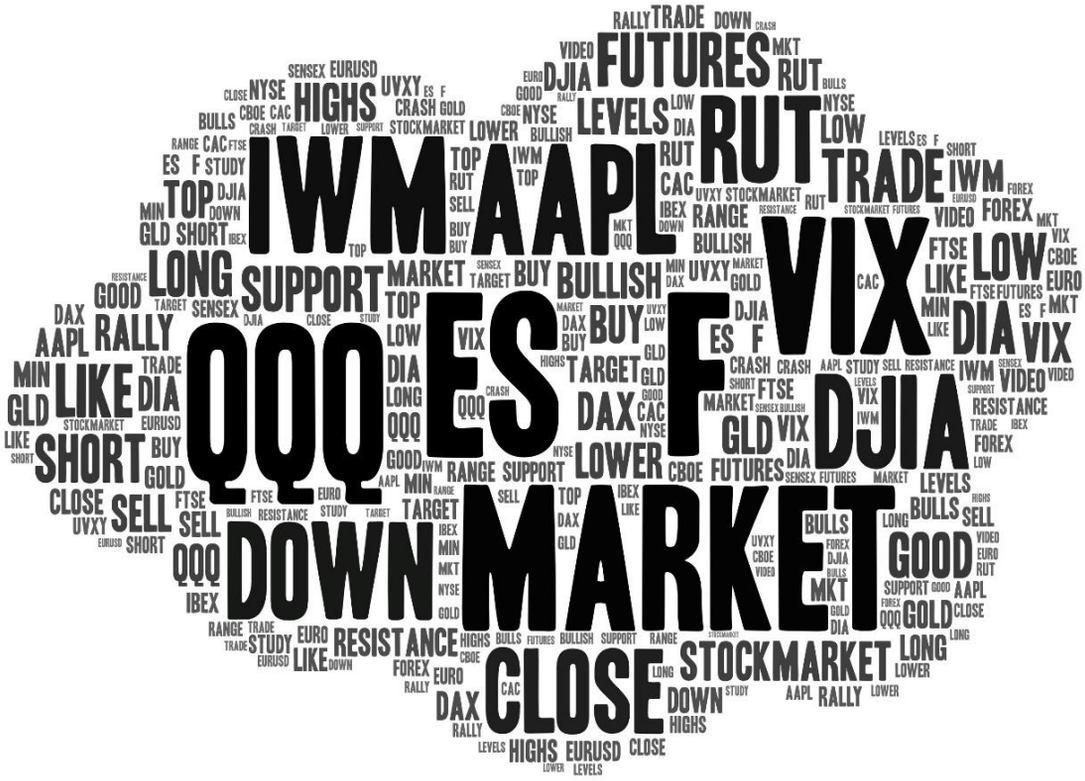

Figure 4. Hot words cloud of the sample period.

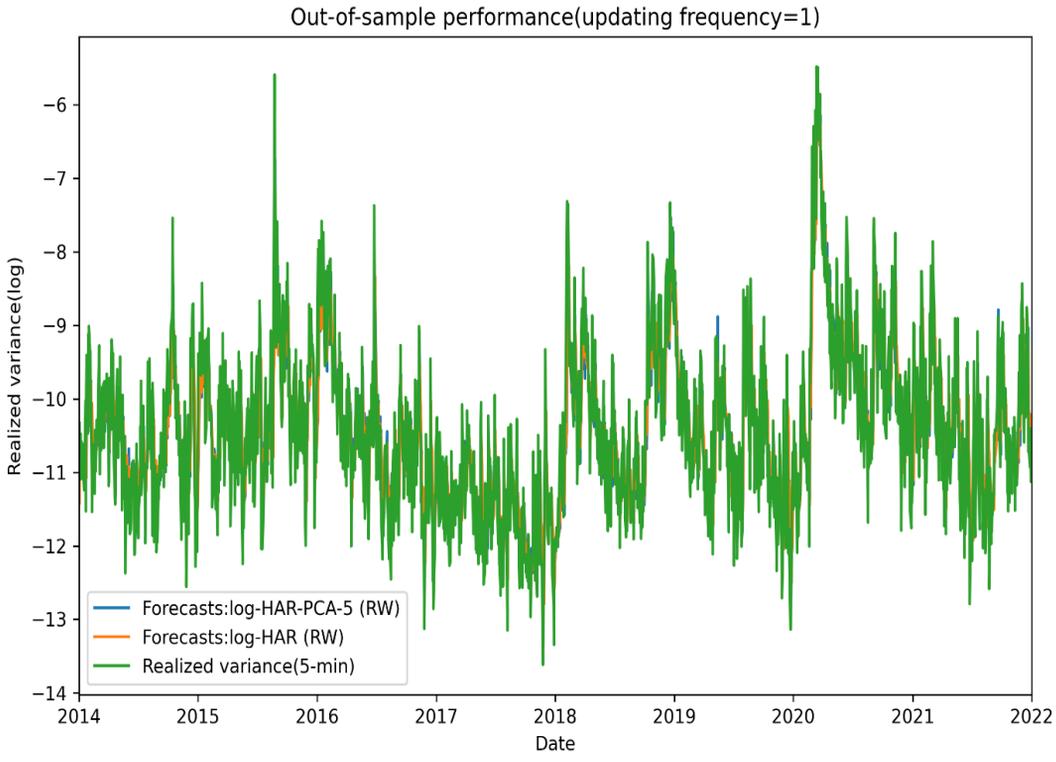

Figure 5. Out-of-sample forecasting (log) comparison of log-HAR-PCA-5 (RW) and log-HAR (RW) with realized variance (5-min)

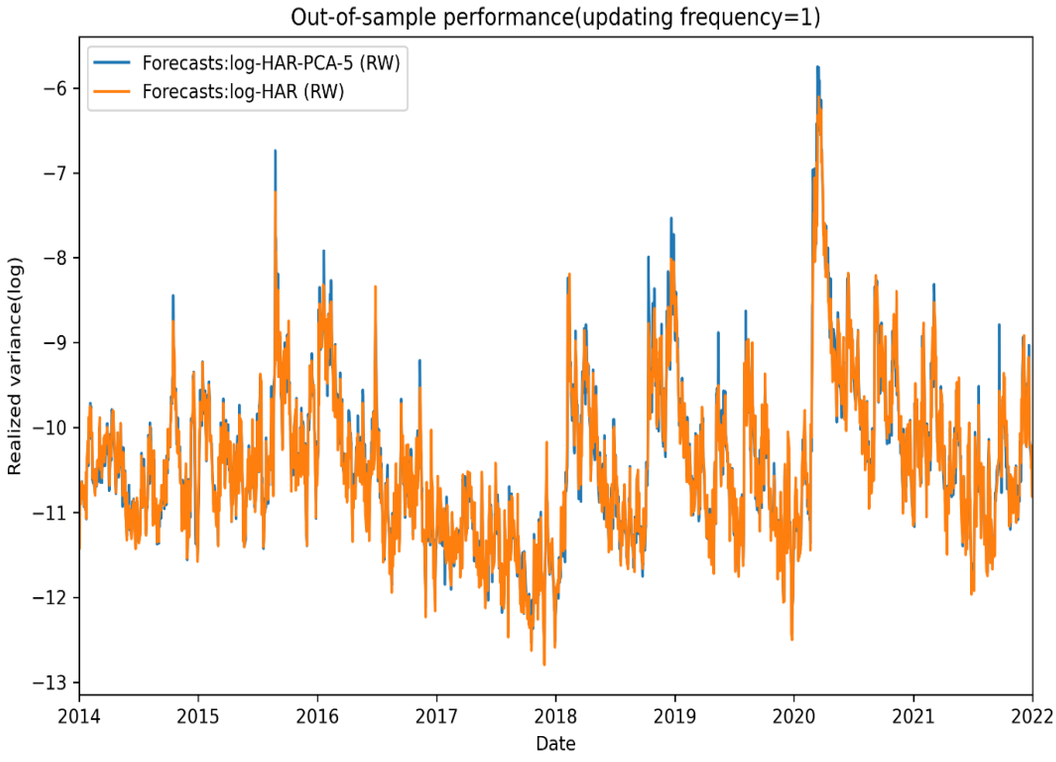

Figure 6. Out-of-sample forecasting (log) comparison of log-HAR-PCA-5 (RW) and log-HAR (RW).

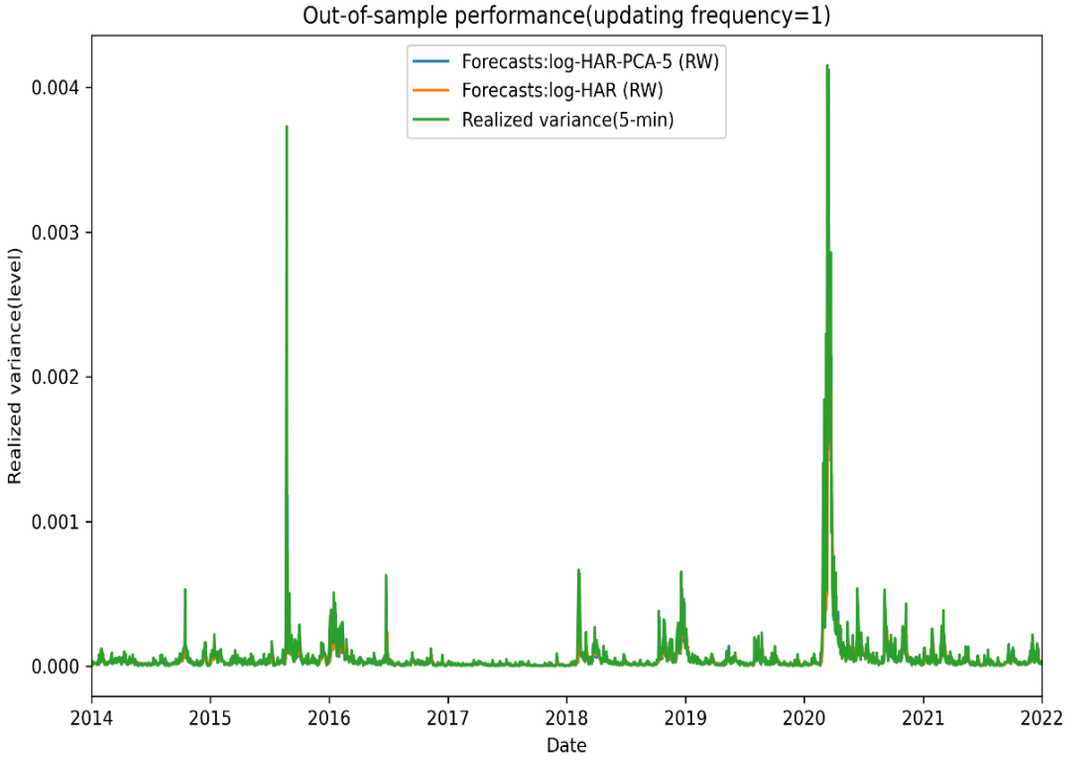

Figure 7. Out-of-sample forecasting comparison of log-HAR-PCA-5 (RW) and log-HAR (RW) with realized variance (5-min)

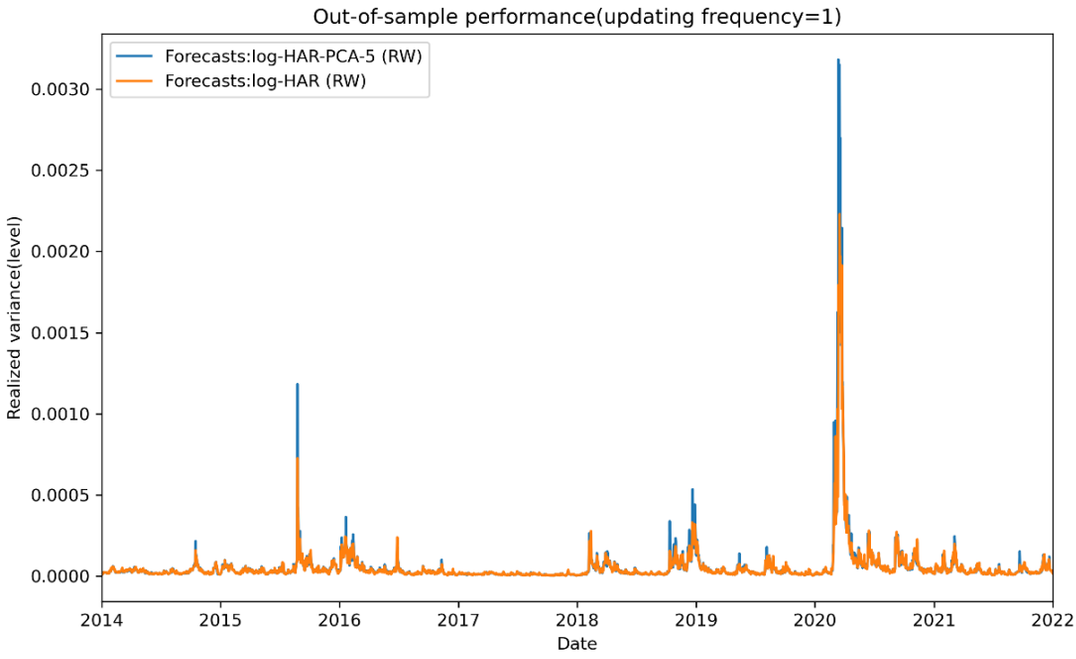

Figure 8. Out-of-sample forecasting comparison of log-HAR-PCA-5 (RW) and log-HAR (RW)