Computational Economics
https://doi.org/10.1007/s10614-025-11042-8

The Sentiment Augmented GARCH-LSTM Hybrid Model for Value-at-Risk Forecasting

Dániel Léber[1] · **Balázs Egyed**[1]

Accepted: 23 June 2025

Abstract
In this paper, we present a new media and social media sentiment-based hybrid GARCH-LSTM model that can more accurately forecast volatility and Value-at-Risk of individual stocks than models proposed by previous studies. Various families of GARCH models and their hybrid extensions have been developed to achieve more accurate conditional volatility forecasts. However, in this paper, we demonstrate that the performance of these models can be significantly enhanced by incorporating different sentiment indicators on external media platforms. We emphasize the nonlinear relationship between variance and sentiment indices and how this relationship can be integrated into various volatility forecasting models. By leveraging the refined predictive capabilities of our extended models, we consolidate them with empirical applicability in the realm of financial risk assessment. Our approach enhances the traditional Value-at-Risk methodology, facilitating a more precise estimation of potential financial losses and offering a robust foundation for strategic risk management decisions. To evaluate our models, we conducted an empirical study on the logarithmic returns of individual stocks comprising the S&P 500 index from 2019 to 2024. In conjunction with the standard Value-at-Risk statistical tests, our study incorporates different loss functions to examine prospective loss magnitudes. Our frameworks demonstrate superior performance in comparison to the traditional GARCH model for a considerable subset of equities, as determined by conventional Value-at-Risk statistical evaluations. Furthermore, comparative analysis indicates that the proposed model ensures the most accurate conditional volatility forecast and its Value-at-Risk estimation achieves the smallest expected loss while satisfying all the statistical tests. These results underscore the superiority of our proposed model in predicting financial risk and volatility efficaciously, although we make further proposals to improve the dissemination of forecasts.

Dániel Léber and Balázs Egyed contributed equally to this work.

Extended author information available on the last page of the article

Published online: 09 July 2025

Keywords Value-at-risk · Twitter sentiments · News sentiments · GARCH-LSTM hybrid model

1 Introduction

Accurate volatility forecasting constitutes a cornerstone in the field of financial risk management and portfolio optimization. The capacity to predict market turbulence with precision is imperative for the formulation of robust risk mitigation strategies and the optimization of asset allocation decisions. In this context, Value-at-Risk has emerged as one of the most popular risk metrics. According to Duffie and Pan (1997) Value-at-Risk (VaR) is a statistical technique used to measure and quantify the level of financial risk of a portfolio over a specific time frame. It estimates the maximum loss that an investment portfolio is likely to suffer, given normal market conditions, over a period of time with a certain degree of confidence. VaR has gained popularity because it is easy to estimate and provides a single, summary measure of market risk for the entire range of financial instruments within a firm's portfolio.

Value-at-Risk has come under a lot of attack, Artzner et al. (1999) highlight that the VaR is not considered a coherent risk measure because of its lack of subadditivity. Subadditivity ensures that the risk of a combined portfolio should not exceed the sum of risks of individual portfolios. According to these authors violations of subadditivity with the VaR can result in financial institutions underestimating risk, which can have implications both for internal risk management and for regulatory capital requirements. As a result of these concerns, Expected Shortfall (ES), which is an alternative risk measure that is coherent and inherently subadditive has been proposed. On the other hand, de Vries et al. (2005) showed that Value-at-Risk can indeed be used as a measure of risk in most practical applications, as it is generally subadditive for common types of assets. His paper finds that concerns regarding VaR's lack of subadditivity are not of significant importance, as violations are rare and mostly confined to assets with extremely fat tails or specific probability levels within return distributions. As such, for typical assets and applications, the VaR remains a reliable and operationally simpler choice for risk measurement compared to other more complicated risk measures.

The Basel Committee on Banking Supervision provides guidelines and standards for banks to manage market risks, including requirements for capital reserves, risk assessment models, and specific reporting practices. As of this writing, the committee provides two possible measures of market risk, Value-at-Risk (VaR) and Expected Shortfall (ES). The move to Expected Shortfall from Value-at-Risk is significant because of the coherence of Expected Shortfall. Nonetheless, one salient challenge that arises with the adoption of Expected Shortfall is the intricacy of its backtesting. Regardless of the choice of the risk metric, backtesting of the VaR remains a key concept in the Basel framework due to its more straightforward application in contrast to backtesting of the Expected Shortfall. The criteria for backtesting necessitate the assessment of each trading desk's daily Value-at-Risk calculations at the 97.5% percent confidence level (Committee, 2013, 2019).

There are different methods for estimating the Value-at-Risk and considering stylized facts in these models is essential because they encapsulate the empirical properties observed in asset returns, which help predict market behavior. In this context, Cont (2001) lists four important stylized facts, volatility clustering, conditional heavy tails, the leverage effect and intermittency. One of the most popular volatility forecast models is General Autoregressive Heteroscedasticity (GARCH) model, which was introduced by Bollerslev (1986). This model can capture the phenomena of volatility clustering and conditional heavy tails in financial time series. Degiannakis et al. (2012) show that the GARCH model has the most satisfactory performance for VaR forecasts before the financial crisis of 2008 for various indices, while during the financial crisis of 2008, the Asymmetric Power GARCH (APARCH) model was found to be more capable of forecasting the VaR efficiently because it is better at capturing extreme negative log returns and the asymmetry of the leverage effect.

On the other hand, for the models to meet these stylized facts, researchers have begun to develop artificial intelligence models, most of which are built on the predictions of GARCH models. Donaldson and Kamstra (1997) reported that the artificial neural network (ANN) model provides the best performance in terms of forecasting volatility because these models are capable of capturing complex, nonlinear relationships in the data that other models cannot. Hu et al. (2020) combined a GARCH model with Long Short-Term (LSTM) and Bidirectional Long Short-Term (BiLSTM) neural networks because these models have memory components and they are useful for capturing time-series characteristics, especially considering sequences of a past event. Kim and Won (2018) presented a comprehensive approach to forecasting stock price index volatility by integrating an LSTM neural network with multiple GARCH-type models such as the Exponential GARCH (EGARCH) model. Their findings also revealed that the hybrid LSTM models significantly outperform existing methodologies, achieving lower prediction errors across various metrics. Kakade et al. (2022) demonstrated that hybrid BiLSTM-GARCH models outperform every conventional model including hybrid GARCH-LSTM models also due to their ability to capture bidirectional dependencies in time series data, providing a more comprehensive understanding of the underlying volatility. García-Medina and Aguayo-Moreno (2024) show another particular advantage of GARCH-LSTM models, they present a powerful tool for predicting the volatility of cryptocurrency portfolios and optimizing portfolio performance according to them. Applying a different approach, Buczynski and Chlebus (2023) introduce a neural network-based GARCH model and show its particular importance in the aspect of Value-at-Risk. Compared with GARCH, the GARCHNet models demonstrate better outcomes during the backtesting. The number of exceedances, which indicates when the actual losses surpass the predicted Value-at-Risk, is lower for GARCHNet models in several cases. GARCHNet models often have lower cost function values from the regulator's perspective, indicating potentially lower costs related to risk management activities. Wang et al. (2023) highlighted the advantages of incorporating multiple features of neural network-based models for forecasting cryptocurrency volatility. The authors identified key variables beyond lagged volatility, these include trading volume, blockchain metrics, Google Trends data, NASDAQ, S&P 500 adjusted close prices and policy uncertainty indices. By utilizing these diverse internal and external determinants,

multivariable neural network models outperform traditional models such as GARCH in terms of forecasting accuracy.

Our study extends the capabilities of conventional GARCH and GARCH-LSTM augmented neural network architectures by attaching sentiment indices drawn from both conventional and social media platforms. Our objective is to enhance the accuracy of the existing models in forecasting market volatility and Value-at-Risk. We analyse the connection between market variances and the sentiment indices, with a particular focus on investigating the nonlinear connection between these variables. Drawing from the foundation established by particularly prior research (Arslan, 2024; Chen & Hu, 2022; Hu et al., 2020; Kakade et al., 2022; Kim & Won, 2018) and our significant results of Granger causality tests, we develop two advanced models: a sentiment index augmented GARCH model and a sentiment index augmented GARCH-LSTM model. These enhanced models are designed to capitalize on the predictive signals embedded within sentiment data, which serve as proxies for investor mood and market perception, thereby providing a more nuanced understanding of market behaviors and future volatility. Grounded in our initial postulations, the integration of sentiment indices is posited not only to encapsulate the characteristically observed stylized facts of fat tails and volatility clustering within financial time series but also to adequately manage the phenomena of intermittency and leverage effects. Additionally, the inclusion of sentiment indices harnesses supplementary information that transcends conventional market data, potentially conferring an enhanced analytical edge in the prognostication of volatility patterns.

We evaluate conventional GARCH, hybrid GARCH-LSTM and our proposed models on empirical data. The forecasted volatilities and corresponding Value-at-Risk estimates derived from the various models under consideration are subjected to rigorous evaluation across the entirety of the S&P 500 index constituents. The comparative analysis of outcomes is evaluated with different test statistics in concert with diverse loss functions, thereby ensuring a comprehensive and multifaceted assessment of the predictive performance of each model. The empirical analysis consists of estimating Value-at-Risk forecasts at the 2.5% significance level one day ahead in a window of 357 test trading days (almost one and a half trading years). The analysis was written and conducted in Python and R.

The paper is organized as follows. In section 2, we present a comprehensive literature review of the importance of sentiment indices and their possible risk management applications. In section 3, we present the data, computational resources and software used. In section 4, we describe the methodology and background for VaR backtesting. In section 5, we show the linear and nonlinear relationships between sentiment indices and social media indices. In section 6, we present the results of the different predictive models, and finally, in section 7 we conclude our results and make further proposals to improve the dissemination of forecasts.

2 Literature Review

One of the critical aspects influencing risk management decisions is the effect of volatility clustering and its consequences. According to Cont (2007) a possible explanation for the phenomenon of volatility clustering is the multi-agent-based model family in which the composition of agents with different trading behaviors changes over time. Lux and Marchesi (1998) differentiate agents according to whether they follow a chartist or a fundamentalist strategy. Chartists predict price trends based on past patterns, while fundamentalists trade based on intrinsic asset values, ignoring short-term fluctuations. Agents probabilistically switch strategies based on realized excess profits, leading to sudden volatility spikes when chartists dominate. This is subsequently followed by market stabilization, a dynamic behavior recognized as another stylized fact known as on-off intermittency. This behavioral switching accounts for not only volatility clustering, heavy tails, and intermittency but also the long-term memory effect observed in volatility. Kirman and Teyssiere (2002) proposed a different multiagent model emphasizing interdependency among agents, leading to herding behavior driven by social influences or rational responses to shared information. Agents adjust their forecasts and trading activities based on current market conditions and past outcomes, contributing to persistent volatility or long memory. The specific rules governing agents' decision updates, such as changing investment thresholds, significantly impact the overall volatility dynamics. According to Banerjee and Green (2015), noise traders, similar to chartists, are irrational agents who interpret noise as informative signals, leading to overreactions or underreactions in asset prices, and contributing to increased volatility and market mispricing. Noise traders are significantly influenced by information from the news and social media such as Twitter, while fundamentalists counterbalance this influence by pushing prices toward their fundamental values. Iqbal et al. (2023) emphasize the psychological biases of noise traders, whose sentiment-driven decisions differ from rational expectations and can create market inefficiencies. All these potential explanations contribute to the dynamics of volatility, which should be considered by risk management.

Many authors have explored the impact of various information sources, including social media and news, on market returns, volume, abnormal returns, and volatility. Antweiler and Frank (2004) investigated the influence of internet stock message boards on financial markets, by analysing over 1.5 million messages from Yahoo! Finance and Raging Bull concerning companies in the Dow Jones Industrial Average and the Dow Jones Internet Index. They discover that high message activity has a small but significant negative predictive effect on stock returns, with transaction costs eroding potential gains. These messages, often posted by noise traders, correlate with increased market volatility and trading volume, suggesting a predictive relationship between message activity and future market volatility. Tetlock (2007) also examines the relationship between media content and stock market movements, focusing on the Wall Street Journal's "Abreast of the Market" column. By quantitatively measuring media pessimism, Tetlock found that high levels of pessimism predict downward pressure on stock prices, followed by a reversion to fundamentals, indicating that media content acts as a proxy for investor sentiment. Additionally, extreme

media pessimism forecasts high market trading volumes and lower market returns lead to greater media pessimism. Tetlock et al. (2008) demonstrated that negative words significantly influence financial analysis more than positive words, as they indicate problems or risks associated with companies. They introduce quantitative measures of negative tone, such as the standardized fraction of negative words, which shows a strong correlation with stock returns and predictive power for various financial attributes. These measures efficiently predict low firm earnings in the upcoming quarter and low stock returns on the following trading day, highlighting the near-immediate impact of negative sentiment on market perception and short-term stock performance. This effect persists even when investors and analysts understand the implications of negative language, as negative words enhance predictions of future earnings beyond commonly accessible quantitative data. The research also reveals a delay in market reactions to negative information, indicating a lag in processing and responding to negative news.

The focus on the fraction of negative words in financial analysis is supported by Baumeister et al. (2001), who find that language contains more words for negative emotions than for positive emotions, emphasizing the greater impact of negative emotions on cognitive processes and judgments of likeability. Negative behaviors are more diagnostic of a person's character, quickly categorizing someone as bad, whereas good behaviors require consistent actions to achieve a positive categorization. Loughran and McDonald (2011) highlight the distinct nature of financial vocabulary compared to conventional language, advocating for a specialized dictionary to accurately measure sentiment in corporate communications. Their research addresses the challenges of word categorization, noting that positive words are often negated in financial contexts, which can cause misclassifications and errors in analysis. For example, phrases such as"did not benefit"can obscure negative information. Another interesting phenomenon that distorts the real meaning of texts is that readers evaluate a company's value based on documents. Considering this phenomenon journalists and authors usually tend to use cautious language, frequently avoiding negative words, instead of qualifying positive words in ways that may not be easily identifiable by different text analysis algorithms. The presence of negative language, however, tends to have a more significant and clear-cut impact. They recommend using their Fin-Neg list, which is better attuned to financial contexts and provides more accurate predictions of corporate performance and market distress than general-purpose negative word lists. This list significantly correlates with financial metrics such as stock returns, volatility, and trading volume, offering a more reliable tool for sentiment analysis in finance and risk management.

Certain authors have already examined the impacts caused by social media, for example, Choi and Varian (2012) explored the use of Google search query data to estimate current economic activity in various industries. They hypothesize that the volume of industry-related Google searches correlates with current performance levels and can anticipate forthcoming data releases. Using Google Trends data, which provides normalized indices of search query volumes from January 1, 2004, they integrate this information into statistical models to estimate real-time economic activity, potentially predicting future events, including volatility. Hamid and Heiden (2015) also investigated the relationship between investor attention, measured by

Google search volumes, and stock market volatility. They enhanced volatility forecasting models by incorporating Google search data and found that it can improve prediction accuracy. Their empirical similarity approach predicts future events by identifying historical market behavior patterns and using similarity-weighted averaging. This method suggests that including investor attention data can refine volatility predictions beyond traditional models, offering innovative insights into economic and financial phenomena.

Deveikyte et al. (2022) explored sentiment analysis for predicting stock market volatility and returns, focusing on the FTSE100 index. Their study confirms a correlation between sentiment in financial news and stock market movements, with negative Twitter sentiment strongly predicting next-day market volatility (with a correlation coefficient of -0.7). By utilizing topic recognition and Latent Dirichlet Allocation, they developed a classifier that achieved 63% accuracy in directional volatility prediction. However, Granger causality tests indicate that sentiment from news and social media does not causally predict changes in the FTSE100 index. Kranefuss and Johnson (2021) also investigated the impact of Twitter sentiment on stock market volatility forecasts using the ARIMA and ARFIMA models. They incorporate sentiment indicators from Loughran McDonald's lexicon and tweet volumes. Their results show that negative Twitter sentiment increases volatility, while positive sentiment decreases it, supporting the leverage effect theory. The volume of Tweets significantly correlates with market volatility, highlighting the growing relevance of social media as an indicator.

Some authors have already established the exploitation of nonlinear relationships between sentiments and financial attributes. Arslan (2024) investigated the challenge of forecasting Bitcoin prices for the following day and the impact of tweets on this price by applying the Long Short-Term Memory (LSTM) model. They utilized Empirical Mode Decomposition (EMD) as a data decomposition technique to discriminate between high and low-frequency components, enabling a comprehensive examination of the attributes inherent in each component. Additionally, they generated a real-world dataset consisting of tweets related to Bitcoin and subsequently processed it for analysis. The proposed scheme demonstrates significant potential for contributing to the rationalization and regulation of the cryptocurrency market, offering guidance to governments and assisting individuals in making profitable investment decisions. Akbiyik et al. (2023) studied Bitcoin volatility forecasting, incorporating data from Twitter as well. They constructed deep-learning models with more than 30 million Bitcoin-related tweets and 15-minute interval price data. Their findings reveal that metadata about Tweet authors, such as influence and follower count, better predict market behavior than Tweet content or simple volume. Temporal convolutional networks outperform classical models, suggesting that identifying influential users is key to predicting Bitcoin volatility.

However, potential sentiment indicators can be constructed using not only Google and Twitter. For instance, Siganos et al. (2014) examine the relationship between investor sentiment on Facebook and stock market returns. Critiquing conventional sentiment measures such as the Michigan Consumer Sentiment survey, they propose using Facebook's Gross National Happiness Index (FGNHI) for daily sentiment analysis across twenty international markets. This method captures daily sentiment

fluctuations and provides natural insights into user sentiment, covering a broader range of countries. Their findings support the hypothesis that positive sentiment biases stock market returns, aligning with behavioral finance theories. In general, a wide range of literature has already demonstrated the strong relationship between various sentiment indices and volatility.

The use of machine learning and deep learning-based models has become a common practice in volatility forecasting. The comparative analysis of Chen and Hu (2022) demonstrates that artificial intelligence models, specifically the ANN and LSTM models, consistently outperform traditional econometric models such as AR and EGARCH in forecasting the volatility of both CSI300 and S&P500 index futures. This performance advantage is evident across multiple loss functions, including mean squared error (MSE), mean absolute error (MAE), and normalized mean squared error (NMSE). Liu (2019) also evaluates the effectiveness of LSTM models in predicting financial volatility, comparing their performance with Support Vector Regression (v-SVR) and the GARCH model. Using data from the S&P500 and AAPL stock indices over a 13-year period, the study demonstrates that LSTM achieves comparable or superior accuracy to v-SVR for large interval predictions and outperforms GARCH for both indices. Key findings reveal that while both LSTM and v-SVR excel over GARCH in minimizing prediction error, LSTM further leverages its deep learning architecture to process raw data effectively, particularly when large datasets are available. The study highlights the scalability and adaptability of LSTM for long-sequence data predictions, albeit with higher computational costs during training. Christensen et al. (2023) examine the effectiveness of different machine learning algorithms in forecasting realized variance for Dow Jones Industrial Average index constituents, comparing them to heterogeneous autoregressive (HAR) models instead of GARCH models. Despite minimal hyperparameter tuning, ML approaches outperform HAR models, particularly at longer forecast horizons. The researchers also attribute this success to ML's ability to capture the long memory of realized variance through higher persistence.

In the comparative analysis of Petrozziello et al. (2022), the effectiveness of LSTM neural networks against traditional econometric models such as Realized GARCH (R-GARCH) and Glosten–Jagannathan–Runkle Multiplicative Error Models (GJR-MEM) is evaluated using data from the NASDAQ-100 and Dow Jones Industrial Average indices. The study uses univariate and multivariate LSTM models to predict realized stock volatility, using data sets that span periods marked by significant economic events, such as the 2007–2008 global financial crisis and the European debt crisis. LSTM models demonstrate superior predictive performance in periods of increased market volatility while maintaining comparable accuracy during tranquil market conditions. In particular, the multivariate LSTM model outperforms others in capturing volatility spillover effects across multiple assets, highlighting its robustness in a multivariate setting. The study of Zhang et al. (2024) explores the efficacy of hybrid models combining the rolling Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) with eight machine learning algorithms to predict high-frequency realized volatility of stock indices and commodity futures. The results show that hybrid models, particularly the CEEMDAN Random Forest model, achieve significant improvements in prediction accuracy compared to stand-

alone machine learning models and traditional econometric approaches, reducing mean square error by up to 52%. We aim to contribute to the literature by incorporating the additional information provided by different sentiments into popular machine-learning methods and emphasizing the nonlinear relationship between variance and sentiment indices.

3 Data Sources, Computational Resources and Software

The data used for the analysis were downloaded exclusively from Bloomberg Terminal. Individual stocks in the S&P 500 stock index for March 2024 are used for the analysis. Individual stocks' adjusted close price, Twitter publication count, Twitter positive sentiment count, Twitter negative sentiment count, News publication count, News positive sentiment, and News negative sentiment count values at a daily frequency from Bloomberg's sentiment analysis tool are downloaded. The data used for analysis have both advantages and disadvantages. The algorithm behind the different sentiment data calculated by Bloomberg and the dictionary used is not public information. However, as mentioned in the literature review, the importance of finding an appropriate dictionary for accurate sentiment analysis has been emphasized in the literature. On the other hand, presumably, a wide range of investors use this feature to set alerts and notifications and most financial institutions have Bloomberg subscriptions. Furthermore, after Twitter became X, the downloading and analysis of Tweets were prohibited and currently there is no other open-source alternative for downloading and analysing tweets. Fortunately, this feature has not been removed from Bloomberg after the company takeover, and subscribers are still provided with up-to-date sentiment information moreover they can download this information at different temporal resolutions for up to five years back.

For the computations in this study, we utilize a high-performance computing environment with an Intel(R) Core(TM) i7-9700K CPU operating at 3.60 GHz, featuring 8 cores and 8 threads and an NVIDIA GeForce RTX 2070 GPU with 8 GB of dedicated RAM. The software tools employed include R Studio for different GARCH model training, as well as Python via the Anaconda distribution and Jupyter Notebook for neural network model development and model backtesting. During the training of neural networks, we leverage the TensorFlow package along with the CUDA and cuDNN toolkits to enable GPU acceleration when available, significantly enhancing the computational efficiency and reducing the training time of our models. With this configuration, the evaluation of nonlinear Granger causality for all stocks in the S&P 500 took approximately five days, whereas the training of the prediction models on the entire dataset took approximately six days. The nonlinear Granger causality analysis was conducted using the Python package implemented by Rosoł et al. (2024), while the expected loss functions implemented by Bayer (2019) were also utilized for backtesting the Value-at-Risk estimates obtained from various models.

4 Methodology

4.1 Realized Volatility

The realized volatility is employed as the target variable during the training of various neural network models, as recommended by Kim and Won (2018). The metric in Formula 1, considering *n* observations provides a precise measure of the historical volatility of an asset's returns over a specified period.

$$\begin{aligned} RV &= \sqrt{\frac{1}{n-1}\sum_{n=1}^{t-1}(r_i - \overline{r})^2} \\ r_i &= \ln\left(\frac{p_{n+1}}{p_n}\right) \end{aligned} \tag{1}$$

where n is the number of observations, p_i is the stock price at the end of the *i*-th interval, r_i the log return at the end of the *i*-th interval, *RV* is the realized volatility and $\overline{r}$ is the mean of r_i.

4.2 Sentiment Indices

One of the pivotal elements in our research is the methodology used to measure media sentiment. Our empirical analysis indicates that employing different sentiment indices results in varying outcomes. Among the various approaches we explore to quantify the negative tone of media, we ultimately adopt the negative sentiment ratio proposed by Tetlock et al. (2008) and subsequently utilized by Kranefuss and Johnson (2021). This decision is based on the compelling rationale presented in their arguments and the significant Granger causalities exhibited by most equities when applying this metric. Nevertheless, we also take into account the positive tone of media with the measure of the positive sentiment ratios, which means that we consider including the indices found in Formula 2 as feature variables. In the analysis, the impact of the negative and positive sentiment indices is examined separately because although these indicators have a negative correlation, they cannot be perfectly expressed because of the inclusion of a nonnegligible amount of neutral sentiments.

$$\begin{aligned} \text{Negative Twitter Sentiment Index}(t) &= -\frac{\text{Number of Negative Tweets}(t)}{\text{Total Number of Tweets}(t)} \\ \text{Positive Twitter Sentiment Index}(t) &= \frac{\text{Number of Positive Tweets}(t)}{\text{Total Number of Tweets}(t)} \\ \text{Negative News Sentiment Index}(t) &= -\frac{\text{Number of Negative News}(t)}{\text{Total Number of News}(t)} \\ \text{Positive News Sentiment Index}(t) &= \frac{\text{Number of Positive News}(t)}{\text{Total Number of News}(t)} \end{aligned} \tag{2}$$

4.3 Measures of Volatility Prediction Performance

To assess and compare the volatility prediction performance of different models, three standard measures of prediction errors are employed, the mean squared error (MSE), the root mean square percentage error (RMSPE) and the mean absolute error (MAE) which are represented in Formula 3.

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \\ \text{RMSPE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{N}|\hat{y}_i - y_i| \end{aligned} \tag{3}$$

where n is the number of observations, y_i is the actual value at observation i, $\hat{y}_i$ is the predicted value of a given model at observation i.

4.4 Granger Causality

Granger causality is used to measure linear causality between returns' variance and different sentiment indices. Granger causality determines whether one time series can help predict another time series. According to Hamilton (2020) in the context of two-time series Y_t and X_t, Granger causality can be tested using the regression model Eq. 4. The Granger causality test involves estimating two models, the first model with only lagged values of Y_t and the other with lagged values of both Y_t and X_t. The F-statistic is subsequently computed to evaluate the joint significance of the coefficients of X_{t-i} in predicting Y_t, grounded in the theoretical framework of Formula 5. The null hypothesis (H_0) is that $\beta_{2i} = 0$ for all *i*, while the alternative hypothesis (H_1) suggests that $\beta_{2i} \neq 0$ for at least one *i*.

$$Y_t = \sum_{i=1}^{p}\beta_{1i}Y_{t-i} + \sum_{i=1}^{q}\beta_{2i}X_{t-i} + \varepsilon_t \tag{4}$$

$$F = \frac{\left(\frac{SSR_r - SSR_{ur}}{m}\right)}{\left(\frac{SSR_{ur}}{n-k}\right)} \tag{5}$$

where Y_t is the dependent variable, X_t is the potential causal variable, ε_t is an error term, p and q are the lag orders, SSR_r is the sum of squared residuals from the restricted model (where the lagged values of X are omitted), SR_{ur} is the sum of squared residuals from the unrestricted model (where the lagged values of X are included), m is the number of restrictions, which is the number of lagged X terms, n

is the number of observations and k is the total number of parameters estimated in the unrestricted model (including both Y and X lags and a constant term).

In our study, we investigate the unidirectional Granger causality between conditional variance and sentiment indices based on Formula 6. We focused exclusively on unidirectional Granger causality because our primary interest is to determine the predictive power of different sentiment indices. In this case, the restricted model includes only lagged values of $\hat{\varepsilon}_t$, while the unrestricted model incorporates lagged values of both $\hat{\varepsilon}_t$ and various lagged values of the specific Sentiment Index_t.

$$\text{Var}(r_t) = \text{E}(\hat{\varepsilon}_t^2 \mid \mathcal{H}_{t-1}) = \alpha_0 + \sum_{i=1}^{p} \alpha_i \hat{\varepsilon}_{t-i}^2 + \sum_{i=1}^{n} \gamma_i \text{Sentiment Index}_{t-i} + u_t \quad (6)$$

where r_t is the log return at time t, $\mathcal{H}_{t-1}$ denotes the history of the process until time t, $\text{Var}(r_t)$ is the conditional variance of r_t, α_0 is a constant term, $\hat{\varepsilon}_{t-i}^2$ is the lagged squared residuals, α_i are the coefficients of the lagged squared residuals, $\text{Sentiment Index}_{t-i}$ is the specific lagged sentiment index with their respective γ_i coefficient and u_t denotes the normally distributed residuals term with zero expected value.

Furthermore, we tested the predictive power of sentiment indices using the nonlinear Granger causality approach proposed by Rosoł et al. (2022). The authors introduce a Python package that measures nonlinear causality with the help of different neural networks. In this case, the unrestricted model is estimated based on a given neural network model represented by Formula 7. The restricted model is estimated as the same neural network without the lagged values of the specific Sentiment Index_t. To evaluate the significance of causality instead of the F-test, the researchers suggest the Wilcoxon signed-rank test in Formula 8. The null hypothesis (H_0) of the statistical test is that the median absolute error for a model based on past values of X is equal to or smaller than that for a model based on past values of X and Y. The alternative hypothesis (H_1) is that the model based on past values of both the X and Y time series has a smaller median absolute error than does the model based only on past values of X.

$$\begin{aligned} \text{Var}(r_t) = f(&\hat{\varepsilon}_{t-1}^2, \hat{\varepsilon}_{t-2}^2, \ldots, \hat{\varepsilon}_{t-p}^2, \\ &\text{Sentiment Index}_{t-1}, \text{Sentiment Index}_{t-2}, \ldots, \text{Sentiment Index}_{t-q}; \theta) \end{aligned} \quad (7)$$

$$\begin{aligned} d_i &= x_{i1} - x_{i2} \\ W^+ &= \sum_{d_i > 0} R(|d_i|) \\ W^- &= \sum_{d_i < 0} R(|d_i|) \\ W &= \min(W^+, W^-) \end{aligned} \quad (8)$$

where f is a specific neural network with a θ vector of hyperparameters, $\mathrm{Var}(r_t)$ is the conditional variance of the r_t log return, $\hat{\varepsilon}^2_{t-i}$ is the lagged squared residual, $\text{Sentiment Index}_{t-i}$ is the lagged sentiment index, p and q are the number of lags of $\hat{\varepsilon}_t$ and Sentiment Index_t respectively, x_i are the individual observations of the two samples, and $R(|d_i|)$ is the rank of the absolute difference $|d_i|$.

4.5 Generalized AutoRegressive Conditional Heteroskedasticity Model (GARCH)

The first pioneering work on volatility forecasting methods was the Autoregressive Conditional Heteroscedasticity (ARCH) model introduced by Engle (1982). The variance and constant mean formulas for the p-th-order ARCH(p) model are given by Formula 9. This model provides a framework for estimating the variance of the error term that is conditional on past information, which is an important aspect of accommodating heteroscedastic behavior in time series data. The parameter p represents the number of lagged squared error terms included in the model.

$$\begin{aligned} r_t &= \mu + \sigma u_t \\ \sigma^2 &= \alpha_0 + \alpha_1 \epsilon^2_{t-1} + \alpha_2 \epsilon^2_{t-2} + \cdots + \alpha_p \epsilon^2_{t-p} \end{aligned} \tag{9}$$

where r_t is the log return at time t, μ is a constant mean, σ^2 is the conditional variance, α_0 is a constant term (and is positive), $\alpha_1, \quad , \alpha_p$ are coefficients to be estimated (which are nonnegative) and u_t and ϵ_t are the error terms that follow a normal distribution with mean zero.

The general extension of the ARCH model was introduced by Bollerslev (1986), who extended the model to incorporate not only past squared errors (which represent past variances) but also past conditional variances. This allows for a more flexible lag structure and can better account for volatility clustering in financial time series. The representation of the conditional variance and constant mean formulas of a GARCH(p, q) model is given by Formula 10. The parameters p and q determine the order of the model, which represents the number of lagged variance terms p and the number of lagged squared error terms q included in the model.

$$\begin{aligned} r_t &= \mu + \sigma u_t \\ \sigma^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2_{t-i} + \sum_{i=1}^{p} \beta_i \sigma^2_{t-i} \end{aligned} \tag{10}$$

where r_t is the log return at time t, μ is a constant mean, σ^2 is the conditional variance at time t, α_0 is a constant term (and is positive), ε^2_{t-i} are the past squared errors, α_i are the coefficients for the past squared errors (which are nonnegative), σ^2_{t-i} are past conditional variances, and β_i are the coefficients for the past conditional variances (which are nonnegative).

4.6 Sentiment Augmented GARCH Model

Engle and Ng (1993) provide a comprehensive examination of how new information is incorporated into volatility estimates, using daily return data from Tokyo Stock Price Index. They introduce the news impact curve to depict the relationship between unexpected returns and resulting predicted volatility, revealing the asymmetric effects of positive and negative news on volatility levels, this phenomenon is often referred to as the leverage effect. Although models such as the Exponential GARCH (EGARCH) model capture much of this asymmetry, they may overestimate the variability of conditional variance, indicating a potential overreaction to past return shocks. Extending GARCH models to include asymmetries caused by new information is crucial.

Based on the findings of previous literature (Degiannakis et al., 2012; Donaldson & Kamstra, 1997; García-Medina & Aguayo-Moreno, 2024; Hu et al., 2020; Kakade et al., 2022; Kim & Won, 2018) and the asymmetric nature of new information's impact on volatility, we propose an extended GARCH model incorporating sentiment indices derived from both traditional and social media sources. The extended model is represented by Formula 11, in which the number of lags associated with the sentiment indices is selected by the Schwarz information criterion. We fit the proposed GARCH model using a hybrid optimization technique, which enhances robustness and accuracy by sequentially employing multiple solvers.

$$
\begin{aligned}
r_t &= \mu + \sigma u_t \\
\sigma^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{i=1}^{r_1} \gamma_i \mathrm{Twitter\_neg}_{t-i} \\
&+ \sum_{i=1}^{r_2} \delta_i \mathrm{Twitter\_pos}_{t-i} + \sum_{i=1}^{r_3} \kappa_i \mathrm{News\_neg}_{t-i} + \sum_{i=1}^{r_4} \lambda_i \mathrm{News\_pos}_{t-i}
\end{aligned}
$$

where r_t is a log return, μ is a constant mean, σ^2 is the conditional variance, α_0 is a constant term, ε_{t-i}^2 are the past squared errors, α_i are the coefficients for the past squared errors, σ_{t-i}^2 are past conditional variances and β_i are the coefficients for the past conditional variances. The additional terms are as follows: $\mathrm{Twitter\_neg}_{t-i}$(negative Twitter sentiment index at time $t-i$), $\mathrm{Twitter\_pos}_{t-i}$(positive Twitter sentiment index at time $t-i$), $\mathrm{News\_neg}_{t-i}$(negative News sentiment index at time $t-i$), and $\mathrm{News\_pos}_{t-i}$(positive News sentiment index at time $t-i$), whose coefficients are γ_i, δ_i, κ_i, and λ_i respectively.

4.7 Multilayer Perceptron

To evaluate nonlinear Granger causality as described in Formula 7, from among several neural network models, we use the multilayer perceptron neural network, which is preferred by Rosoł et al. (2022). According to Rumelhart et al. (1986) a multilayer perceptron (MLP) is a class of feedforward artificial neural networks that consists of multiple layers of nodes, which are typically interconnected in a feedforward way.

Each node in the network, often called a neuron, is a computational unit that takes input from the previous layer, applies a transformation, and provides output to the next layer. The key components of an MLP include the input layer, hidden layers, and the output layer. The input x_i to the *i*-th neuron is a weighted sum of outputs from neurons in the previous layer, as shown in Formula 11. The output y_i of the *i*-th neuron is determined by applying a linear or nonlinear activation function *f* to the input x_i, as shown in Formula 12. Common choices for *f* include the sigmoid function, hyperbolic tangent function (tanh) or rectified linear unit function (ReLU). The error (*E*) in the output is often computed using a cost function and a few examples are shown in Formula 3. To update the weights, backpropagation is used which involves computing the derivative of the error cost function with respect to each *w* weight using the chain rule shown in the Formula 13 and updating the weights according to Formula 14, where η is a learning rate parameter. Adaptive Moment Estimation (ADAM) is an optimization algorithm that we use during neural network training, that provides an adaptive learning rate for each parameter. It leverages the first and second moments of the gradients to dynamically adjust the learning rate, resulting in faster convergence and improved performance on large datasets or models with many parameters (Kingma & Ba, 2015). Figure 1 shows an example of a schematic of a multilayer perceptron architecture with 4 *x* feature variables, one hidden layer with 3 *n* neurons and an output layer with one neuron. Every neuron's input in Fig. 1 is the w_i-weighted sum of the previous units, and its output is the value of the sum evaluated by a specific activation function.

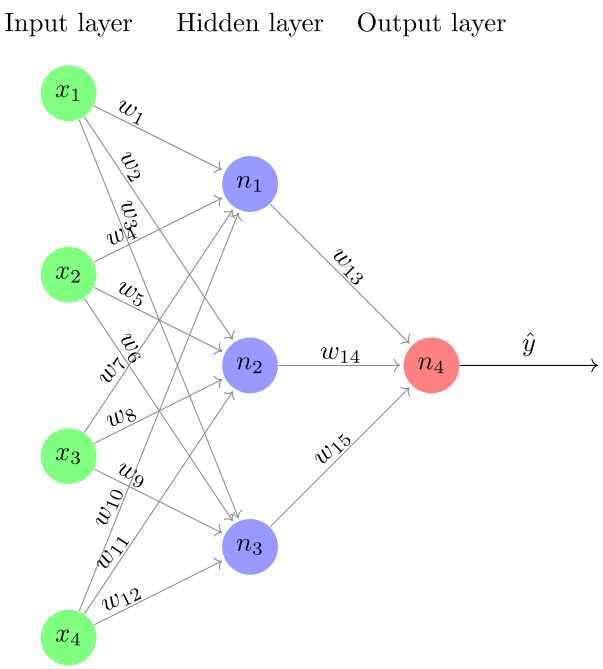

Fig. 1 Schematic architecture of an arbitrary neural network with 4 inputs, one hidden layer with 3 neurons and an output layer with 1 neuron

$$x_i = \sum_j w_{ij} y_j + b_i \tag{11}$$

$$y_i = f(x_i) = f\left(\sum_j w_{ij} y_j + b_i\right) \tag{12}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial x_i} \cdot \tag{13}$$

$$w_{ij}(t+1) = w_{ij}(t) - \eta \cdot \frac{\partial E}{\partial w_{ij}(t)} \tag{14}$$

where w_{ij} represents the weight from the *j*-th neuron in the previous layer to the *i*-th neuron in the current layer, y_j is the output from the *j*-th neuron in the previous layer, b_i is the bias term for the *i*-th neuron, f is a linear or nonlinear activation function, E is the cost function and η is a learning rate parameter.

When applying neural networks, one of the most important considerations is the issue of overfitting. Overfitting occurs when a model learns the training data too precisely, including noise and random fluctuations, mistaking them for genuine patterns. This typically leads to poor generalization to new, unseen data which is a possible consequence of poor predictive power. One of the possible solutions is proposed by Srivastava et al. (2014), a technique known as dropout, which involves randomly removing units from the neural network during training. This approach prevents the coadaptation of features and compels the network to learn more robust and generalizable features. Another useful technique was presented by Morgan and Bourlard (1989), who proposed stopping the training process before the network has fully converged to the minimum training loss, based on the performance on a validation set. The training is halted when the network's error on the validation set starts to increase, indicating that the model is beginning to overfit the training data, rather than learning general patterns. The benefit of so-called early stopping is that it helps to maintain a good generalization performance of the neural network on new, unseen data. By observing the network's performance on an independent test set and stopping the training at the right time, one can reduce the sensitivity to the choice of network size and the amount of training, thus helping to avoid the waste of computational resources and potentially achieving more robust performance.

4.8 Long Short-Term Memory (LSTM)

Rather than combining GARCH models with a simple multilayer perceptron approach, it is advisable to use methods that account for the memory of conditional volatility. According to the literature, a more effective combination involves the Long Short-Term Memory (LSTM) recurrent network family (Chen & Hu, 2022; García-Medina & Aguayo-Moreno, 2024; Hu et al., 2020; Kakade et al., 2022; Kim & Won, 2018; Liu, 2019; Lu et al., 2016; Petrozziello et al., 2022). Recurrent neural networks

(RNN) are a type of neural network architecture that is particularly well-suited for processing sequences of data such as time series. The key feature of RNNs is that they have connections that loop back on themselves, allowing them to maintain information about previous inputs within their hidden state. This makes them powerful for tasks where context from earlier in the sequence is important for making predictions. LSTM is a specialized form of RNN proposed by Rumelhart et al. (1986) and is designed to overcome the limitations of traditional RNNs, particularly the difficulty in learning dependencies between events that occur at vastly different points in time. It achieves this through the use of a complex cell structure that includes multiple gates controlling the flow of information. The formal form of one LSTM cell is represented in Formula 15 and a visual representation of the formulas can be found in the schematic Fig. 2. Each formula corresponds to one step in the operation of the LSTM cell at a given time step *t*.

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{15}$$

where f_t is the forget gate activation vector, i_t is the input/update gate activation vector, o_t is the output gate activation vector, $\tilde{C}_t$ is the cell input activation vector, C_t is the cell state vector, h_t is the hidden state vector/output vector of the LSTM unit, σ is the sigmoid function, $\tanh$ is the hyperbolic tangent function, W are the weight matrices for each of the gates, b are the bias terms for each of the gates and $[h_{t-1}, x_t]$

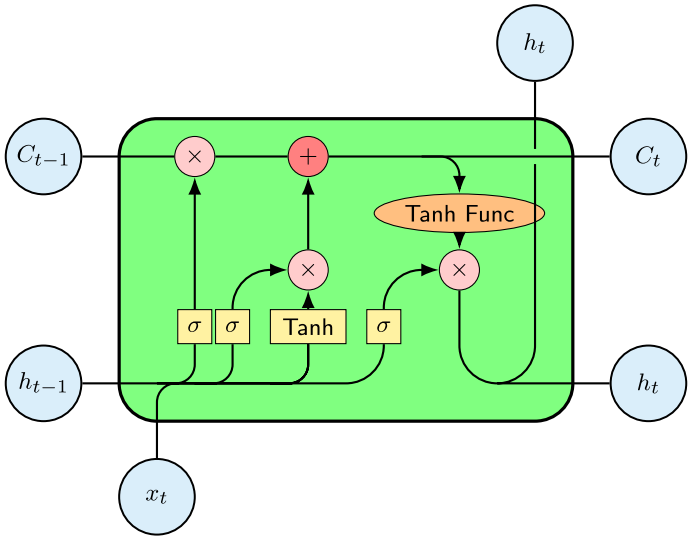

Fig. 2 The schematic architecture of one LSTM cell

indicates the concatenation of the previous hidden state and the current input. The $*$ operation represents the Hadamard product.

Schuster and Paliwal (1997) advanced the LSTM cell-based neural networks, enhancing their performance on sequence prediction problems. In a standard LSTM network, information flows in a single direction through the sequence from past to future. However, in situations where the full context of the sequence is required to achieve the most accurate predictions, Bidirectional long short-term memory (BiLSTM) is especially advantageous. One BiLSTM layer consists of two LSTMs, one processes the sequence from start to end, while the other processes it from end to start. This setup allows the network o have both past and future context at each point in the sequence, potentially capturing patterns that may be missed by unidirectional LSTM models. Kakade et al. (2022) and Hu et al. (2020) demonstrated the superiority of GARCH-based BiLSTM models over traditional models, for this reason in our model, we also consider a BiLSTM layer.

4.9 Value-at-Risk Estimation

The Value-at-Risk (VaR) defines the potential loss in value of a portfolio over a period of time, given normal market conditions with a given confidence interval. According to the definition the VaR can be expressed as Formula 16 (Duffie & Pan, 1997). If we assume that higher-order moments do not play a role and the distribution of returns can be considered normal, then the Value-at-Risk can be estimated using the delta-normal method as shown in Formula 17.

$$\mathrm{VaR}_\alpha = F^{-1}(1-\alpha) \tag{16}$$

$$\mathrm{VaR}_\alpha = Z_\alpha \cdot \sigma \cdot \sqrt{\Delta t} \tag{17}$$

where VaR_α represents the Value-at-Risk at the α confidence level, F^{-1} is the inverse cumulative distribution function, Z_α is the quantile of the standard normal distribution corresponding to the α confidence level, σ is the volatility of the returns and Δt is the given time horizon.

4.10 Value-at-Risk Backtesting

One type of Value-at-Risk statistical test is the unconditional coverage test, we exclusively focus on the Kupiec test which is commonly used to assess the adequacy of a Value-at-Risk (VaR) model. It compares the actual number of exceedances of the VaR against the expected number under the assumed distribution. The likelihood ratio test is used to compare the goodness-of-fit of two nested models for estimating the proportion of failure (POF) in a binary outcome scenario (0 for success, 1 for failure). The test statistic supposes that two models are available for estimating the proportion of failure, the null model M_0 and the alternative model M_1. In the null model M_0, where we assume a fixed proportion of failure, represented by π_0, the corresponding likelihood function is given by Formula 18. In the case of the alternative model M_1, we estimate the proportion of failure as a parameter, denoted by π_1,

whose likelihood function is given by Formula 19. The likelihood ratio test statistic compares the two models according to Formula 20. Under certain regularity conditions, the likelihood ratio test statistic LR_{POF} asymptotically follows a chi-squared distribution with one degree of freedom. According to the null hypothesis (H_0) the proportion of failure is equal to π_0, while the alternative hypothesis (H_1) suggests that the proportion of failure is different from π_0.

$$\mathcal{L}(M_0) = \prod_{i=1}^{n} (\pi_0)^{y_i} (1 - \pi_0)^{1-y_i} \tag{18}$$

$$\mathcal{L}(M_1) = \prod_{i=1}^{n} (\pi_1)^{y_i} (1 - \pi_1)^{1-y_i} \tag{19}$$

$$LR_{POF} = -2 \ln \left(\frac{\mathcal{L}(M_0)}{\mathcal{L}(M_1)} \right) \tag{20}$$

where y_i is the binary outcome for observation *i* (0 for success, 1 for failure), π_0 represents the proportion of failures in the M_0 model, π_1 represents the proportion of failures in the M_1 model and *n* is the total number of observations.

Another type of Value-at-Risk test is the independence test, in this case, we focus on the Christoffersen and Pelletier (2004) duration test which examines the time between the *i*-th and $(i-1)$-th violations. The correctly specified Value-at-Risk model would imply independently and identically distributed durations. The durations under the null hypothesis should follow an exponential distribution with the probability density function expressed by Formula 21 and a constant hazard function. To test the independence of these durations, we can compare the observed distribution against a Weibull distribution as described by Formula 22. The null hypothesis (H_0) of the statistical test is that $b = 1$, therefore the durations are memoryless. The alternative hypothesis (H_1) is that $(b \neq 1)$, implies that durations are not memoryless.

$$f_{\exp}(D) = pe^{-pD} \tag{21}$$

$$f_W(D) = ab^b D^{b-1} \exp(-bD^b) \tag{22}$$

where $D_i = t_i - t_{i-1}$ is the duration, in which t_i is the time of the *i*-th violation, *p* is the rate parameter of the exponential distribution and *a*, *b* are the scale and shape parameters of the Weibull distribution respectively.

According to Campbell (2005) an accurate VaR measure must satisfy both the independence and unconditional coverage properties. Tests that simultaneously evaluate these properties are capable of identifying deficiencies in Value-at-Risk measures. Kupiec's unconditional coverage test and Christoffersen and Pelletier's duration test can be combined into one statistical test expressed by Formula 23. The combined likelihood ratio statistic follows a chi-squared distribution with 2 degrees of freedom.

$$LR_{\text{Joint}} = LR_{\text{POF}} + LR_{\text{Duration}} \tag{23}$$

where LR_{POF} is the likelihood ratio statistic for Kupiec's Proportion of Failures test and LR_{duration} is the likelihood ratio statistic for Christoffersen and Pelletier's duration-based test.

The Value-at-Risk is usually also tested by different loss functions. These functions inherently adopt a negative orientation, attributing higher values to instances of model failure, thus penalizing inaccurate predictions more severely. The VaR model that yields the minimal loss value is deemed superior, as it demonstrates a greater efficacy in managing and mitigating risk. This evaluative process ensures that the chosen model aligns closely with the risk preferences and operational objectives of the risk managers. Lopez et al. (1999) formalized loss functions and proposed a quadratic loss function known as the magnitude loss function. Lopez's binomial loss is presented in the Formula 24 and the quadratic loss function can be expressed as an expected value of the $C_{m,t}$ cost function.

$$C_{m,t+1} = \begin{cases} 1 + (r_{t+1} - \text{VaR}_{m,t+1})^2 & \text{if } r_{t+1} < \text{VaR}_{m,t+1} \\ 0 & \text{if } r_{t+1} \geq \text{VaR}_{m,t+1} \end{cases} \tag{24}$$

where $C_{m,t+1}$ is the cost function for model m at time $t+1$, r_{t+1} is the actual return at time $t+1$ and $\text{VaR}_{m,t+1}$ is the predicted value of Value-at-Risk for model m at time $t+1$.

According to Sarma et al. (2003), the loss function of a firm is designed to capture the economic costs associated with the failures of a Value-at-Risk (VaR) model from a firm's perspective. This includes not only direct losses when the actual portfolio returns fall below the VaR threshold but also the opportunity cost of holding excessive capital. The function emphasizes the trade-off between safety and capital efficiency, penalizing both model failures and excessive capital allocation. The binomial loss of the firms can be expressed by Formula 25 and the firm loss is an expected value of the l_t loss functions.

$$l_{t+1} = \begin{cases} (r_{t+1} - \text{VaR}_{t+1})^2 & \text{if } r_t t + 1 < \text{VaR}_{t+1} \\ -\alpha \text{VaR}_{t+1} & \text{otherwise} \end{cases} \tag{25}$$

where l_{t+1} is the loss at time $t+1$, r_{t+1} is the actual return at time $t+1$, VaR_{t+1} is the predicted Value-at-Risk at time $t+1$ and α is the opportunity cost of capital.

According to González-Rivera et al. (2004), the smoothed Value-at-Risk (VaR) loss function provides an advanced method for evaluating VaR models. Unlike traditional nondifferentiable loss functions, it offers a smooth approximation that facilitates optimization and statistical inference. In Formula 26 the smooth approximation function $m_\delta(a,b)$ makes the loss function differentiable, which is crucial for the application of gradient-based optimization techniques.

$$\begin{gathered} \widetilde{Q} = P^{-1} \sum_{t=R}^{T} \left(\alpha - m_\delta(r_{t+1}, \text{VaR}_{t+1}^{\alpha})\right) (r_{t+1} - \text{VaR}_{t+1}^{\alpha}) \\ m_\delta(a,b) = [1 + \exp\{\delta(a-b)\}]^{-1} \end{gathered} \tag{26}$$

where α is the quantile level for the Value-at-Risk (VaR) calculation, r_{t+1} is the actual return or value at time $t+1$, $\mathrm{VaR}_{t+1}^{\alpha}$ is the predicted Value-at-Risk at time $t+1$ for the given quantile level α, P is the prediction period, R is the estimation period, T is the total number of periods and δ is the smoothness parameter for the function $m_{\delta}(a,b)$, which controls the approximation of the indicator function.

5 Predictive Power of Sentiment Indices

In this section, we utilize both linear and nonlinear Granger causality statistical tests at different significance levels to assess the predictive power of various sentiment indices on the variance of firms listed on the S&P 500. It is essential to note that the focus here is on predictive power rather than strict causality, highlighting the capability of sentiment indices to forecast future stock variances based on historical observations. To ensure the robustness of our findings, we leverage the entire available period of historical data (5 years) for each firm in the S&P 500.

5.1 Linear Granger Causality

In the context of assessing the predictive power of different sentiment indices on stocks, we begin by ensuring that both the firm's variance and the related sentiment index are stationary. This is a critical step as nonstationary data can lead to spurious results in Granger causality tests. To determine stationarity, we utilize the Augmented Dickey-Fuller (ADF) test (Hamilton, 2020). In Table 1 we present the results of the ADF tests, which indicate the proportion of companies in the S&P 500 whose variance and standardized sentiment indices are together stationary at the 5% significance level.

The linear Granger causality test is conducted only in cases where both the variance and the given standardized sentiment index indicators are stationary at the 5% significance level. The statistical tests are performed between the variance and the different standardized sentiment indices with lags ranging from 1 to 10. The proportions of companies for which the linear Granger causality between different sentiment indices and variance is significant at different levels are presented in Tables 2, 3, and 4. The linear Granger causality tests reveal significantly better results in terms of the relationship between the variance of stock returns and the negative Twitter sentiment index than other sentiment indices. These results for negative tone align with those of Tetlock et al. (2008), highlighting the predictive power of negative sentiments for volatility and variance. The analysis reveals that the variance of more than one-third of the firms in the S&P 500 can be effectively predicted at the 10% significance level using the standardized negative Twitter sentiment index from the

Table 1 The proportion of companies in the S&P 500 whose variance and standardized sentiment indices are together stationary at the 5% significance level

Sentiment Index	Propotion of Stocks
Negative Twitter Sentiment Index	91%
Positive Twitter Sentiment Index	90%
Negative News Sentiment Index	66%
Positive News Sentiment Index	92%

Table 2 Proportions of observed companies for which the linear Granger causality between sentiment indices and variance are significant at the 1% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	9%	3%	2%	1%
2	10%	3%	2%	2%
3	10%	3%	2%	1%
4	10%	4%	3%	1%
5	7%	4%	2%	1%
6	7%	4%	3%	2%
7	7%	4%	4%	3%
8	8%	3%	3%	3%
9	7%	3%	4%	3%
10	6%	3%	4%	3%

Table 3 Proportions of observed companies for which the linear Granger causality between sentiment indices and variance are significant at the 5% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	27%	9%	6%	6%
2	21%	10%	5%	5%
3	20%	8%	6%	3%
4	18%	9%	5%	3%
5	17%	8%	6%	3%
6	16%	6%	7%	3%
7	15%	6%	7%	4%
8	16%	5%	6%	4%
9	15%	5%	5%	3%
10	14%	6%	5%	4%

Table 4 Proportions of observed companies for which the linear Granger causality between sentiment indices and variance are significant at the 10% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	35%	16%	11%	13%
2	30%	14%	8%	9%
3	26%	13%	8%	7%
4	24%	12%	8%	6%
5	24%	11%	8%	5%
6	21%	10%	8%	5%
7	20%	8%	7%	5%
8	20%	8%	7%	5%
9	20%	8%	8%	5%
10	21%	8%	8%	5%

preceding period. However, following the first lag, this proportion begins to rapidly decrease and the proportion of the preceding period at 1% significance is only 10%. Nonetheless, the proportion of significant results at the 5% and 10% significance levels caused by negative Twitter sentiment remains surprisingly high, even at the 10th lag. The proportion of significant results observed with other sentiment indices is considered notable only at the 10% significance level.

5.2 Nonlinear Granger Causality

To determine whether there is a nonlinear relationship between the variance and different sentiment indices, we evaluate the linear Granger causality approach proposed by Rosoł et al. (2022). The restricted and unrestricted models of the statistical test are specified as described in Formula 7, where following the authors' recommendation, we utilize a multilayer perceptron to approximate the nonlinear function. The specifications of the multilayer perceptron and the training process are summarized in Table 5. To achieve the best results, the training is conducted using multiple learning rates. Furthermore, the training process is repeated five times with different initial weights in each case. The model that achieves the smallest Residual Sum of Squares (RSS) on the validation set, utilizing the corresponding trained weights, is employed to make predictions on the test set.

To investigate the nonlinear relationship between variance and various sentiment indices, we consider the lags of these sentiment indices from 1 to 10. The proportions of companies for which the nonlinear Granger causality between various sentiment indices and variance is significant at the 1%, 5%, and 10% significance levels are presented in Tables 6, 7, and 8. The results indicate that, at all conventional significance levels (1%, 5%, and 10%), the variance of a larger proportion of the examined companies can be predicted by the given sentiment index compared to the result of linear Granger causality. The hierarchical structure of the significant causality proportions between different sentiment indices remains consistent. Specifically, at every significance level the highest proportion of significant Granger causality is observed between the negative Twitter sentiment index and the variance. This is followed by the proportions of the positive Twitter sentiment index, negative News sentiment index, and finally the positive News sentiment index. Lowering the significance level only slightly reduces the proportion of significant results and at a given significance

Table 5 Multilayer Perceptron's Hyperparamaters of Non-linear Granger Causality Test

Parameters	Value
Number of Hidden Layers	2
Number of Neurons at Each Layer	100
Activation Functions at Each Layer	Rectified Linear Unit (ReLU)
Epochs	30
Learning Rates	0.0001, 0.00001
Batch Size	32
Dropout Rate after Every Hidden Layer	5%
Validation Ratio	20%
Test Ratio	30%
Loss Function	MSE

Table 6 Proportions of observed companies for which the nonlinear Granger causality between sentiment indices and variance are significant at the 1% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	81%	79%	32%	28%
2	78%	76%	32%	24%
3	74%	72%	26%	26%
4	69%	68%	20%	18%
5	69%	64%	19%	18%
6	65%	61%	14%	15%
7	64%	59%	12%	13%
8	60%	58%	9%	9%
9	58%	55%	9%	12%
10	54%	54%	9%	8%

Table 7 Proportions of observed companies for which the nonlinear Granger causality between sentiment indices and variance are significant at the 5% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	83%	81%	39%	37%
2	79%	79%	37%	31%
3	78%	75%	32%	31%
4	72%	74%	26%	26%
5	73%	68%	23%	23%
6	70%	65%	19%	20%
7	68%	64%	18%	18%
8	67%	62%	15%	15%
9	64%	58%	14%	16%
10	62%	59%	13%	11%

Table 8 Proportions of observed companies for which the nonlinear Granger causality between sentiment indices and variance are significant at the 10% significance level across different lags

Lag	Negative Twitter	Positive Twitter	Negative News	Positive News
	Sentiment Index	Sentiment Index	Sentiment Index	Sentiment Index
1	84%	83%	42%	41%
2	81%	79%	40%	35%
3	78%	77%	36%	33%
4	74%	76%	30%	31%
5	75%	72%	26%	27%
6	73%	67%	23%	24%
7	69%	66%	22%	23%
8	67%	64%	19%	18%
9	67%	61%	18%	20%
10	65%	63%	16%	15%

level increasing the number of lags noticeably reduces the proportion of significant results. The first lag of the given Twitter sentiment indices offers predictive insights of variances for more than three-quarters of the observed companies. In contrast, the first lag of the News sentiment indices provides useful information for predicting the variance in less than half of the companies examined. However, this proportion remains significant, indicating that for nearly half of the companies, news sentiment indices can offer valuable predictive insights. Therefore, despite being less effective than Twitter sentiment indices, the predictive power of news sentiment indices should not be neglected. A similar phenomenon is observed when considering the last examined lag. In this case, the Twitter sentiment indices exhibit significant predictive power for more than half of the companies studied. In contrast, the corresponding predictive capability of the News sentiment indices becomes negligible. These results suggest that in a nonlinear model, indices calculated from Twitter sentiment contain significantly more information for predicting stock variance than indices derived from news sentiment.

6 Performance Evaluation and Backtesting of Predictive Models

In this chapter, we present the hyperparameters of our predictive models and the configuration used for training and backtesting. Additionally, we present the key observations on randomly selected stocks, followed by a comprehensive presentation of the results for all stocks in the S&P 500 index.

6.1 Hyperparamaters of Neural Networks

We use the one-day ahead expanding rolling window technique to forecast 1-day Value-at-Risk values at the 2.5% significance level using four models. The first model is the traditional GARCH(1,1) model presented in Formula 10, referred to as *GARCH*. The second model is an extended GARCH model presented in Formula 11, referred to as $GARCH\_X$. The third and fourth models are LSTM neural network models, both of which are based on GARCH(1,1) forecasts. These models differ solely in their input variables. We refer to the first neural network model as the *NN* model which consists solely of contemporaneous and lagged volatility forecasts from the GARCH(1,1) model. The second neural network model is referred to as the $NN\_Sentiment\_Index$ model, whose input is augmented with different lagged sentiment index values. The look-back period of both models is one month (21 trading days). The target variable in both cases is the realized volatility computed from 10 days. First, we train the GARCH(1,1) model exclusively on the first three months (63 trading days) of the entire 5-year time series. Second, the LSTM neural networks are trained on 70% of the remaining data. The training is repeated three times with different initial weights in each case. The best model is selected based on the lowest RMSPE on the validation set, which comprises 30% of the training set. Finally, we backtest the predictive capabilities of the different models on the test dataset, which is the commonly applied 30% of the observed dataset (almost one and a half trading years, 357 trading days). According to Campbell (2005), the minimum period exam-

ined during the backtest is one year because the possibility that no violations occur during a short period such as one year is not trivial. Both of the neural networks' first hidden layers are BiLSTM layers, followed by three LSTM hidden layers. During hyperparameter tuning, we observe that the application of multiple BiLSTM layers negatively impacts the performance of both models. The selected hyperparameters of the two models are presented in Table 9. To avoid overfitting, we employ minibatch training, dropout between the hidden layers, and early stopping.

6.2 Results on Randomly Selected Stocks

The following analysis illustrates the performance differences among various models using three randomly selected stocks characterized by high volatility and intermittency. The selected stocks are International Alphabet Inc. (GOOG), Business Machines Corporation (IBM) and Tesla Inc. (TSLA). The results of the Value-at-Risk (VaR) at the 2.5% significance level forecasted by different models over the test period are visually presented in Figs. 3, 4, and 5. The forecasts generated by the *GARCH* and $GARCH\_X$ models exhibit significant differences, attributed to the additional information considered in the $GARCH\_X$ model. Compared to the *GARCH* model, the $GARCH\_X$ model is more responsive to unforeseen changes in returns, and its absolute magnitudes are occasionally greater. Furthermore, in many instances, the Value-at-Risk levels forecasted by the $GARCH\_X$ model are generally lower than those forecasted by the *GARCH* model. This endeavors to address the underestimation of Value-at-Risk observed in the GARCH model family as noted by Lee and Su (2012).

The neural network models, $NN\_Sentiment\_Index$ and *NN* exhibit heightened sensitivity to immediate changes, indicating their ability to swiftly respond to abrupt market fluctuations. The *NN* model, which relies exclusively on lagged conditional variances predicted by the *GARCH* model, attempts to react more promptly than the simple *GARCH* model. However, the *NN* model is unable to capture all changes, for instance, there are certain changes to which only the $GARCH\_X$ model can respond promptly, due to the additional information it incorporates. The sentiment-augmented neural network $NN\_Sentiment\_Index$ leverages the additional information utilized by the $GARCH\_X$ model to better capture and react to sudden and intense market changes. The absolute magnitudes of the VaR forecasts produced by

Table 9 The hyperparameters of neural networks for volatility forecasting

Parameter	Value
Hidden layers	BiLSTM, LSTM, LSTM, LSTM
Number of LSTM cells in hidden layers	128, 64, 32, 16
Activation functions of neurons	Hyperbolic Tangent
Dropout ratio between hidden layers	0.1
Learning rate	0.001
Epochs	200
Early stopping patience	30 Epochs
Batch Size	32
Loss function	RMSPE

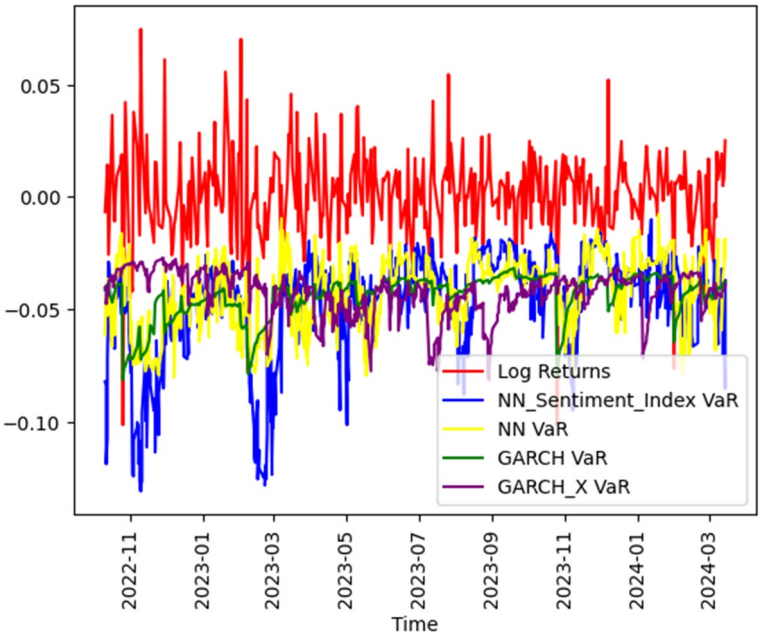

Fig. 3 Forecasts of the 2.5% Value-at-Risk for Alphabet Inc. stock in the test dataset as predicted by multiple models

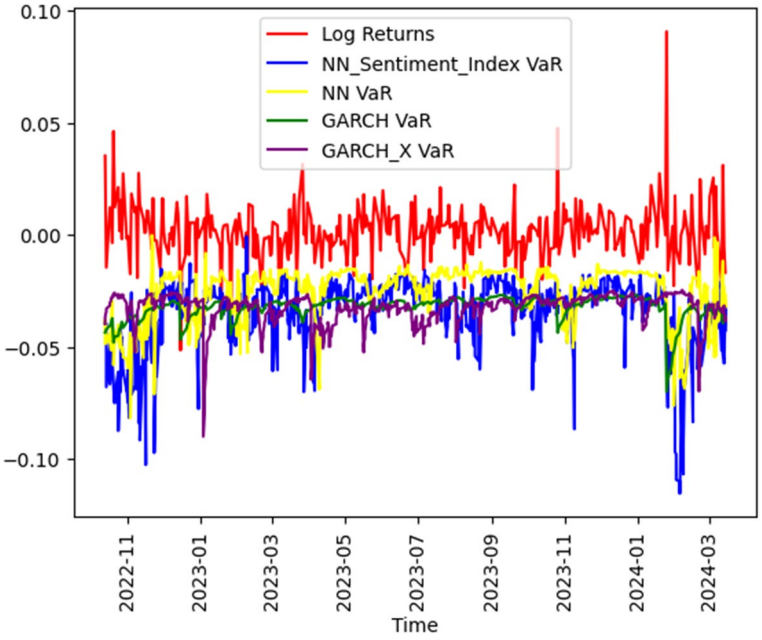

Fig. 4 Forecasts of the 2.5% Value-at-Risk for Business Machines Corporation stock in the test dataset as predicted by multiple models

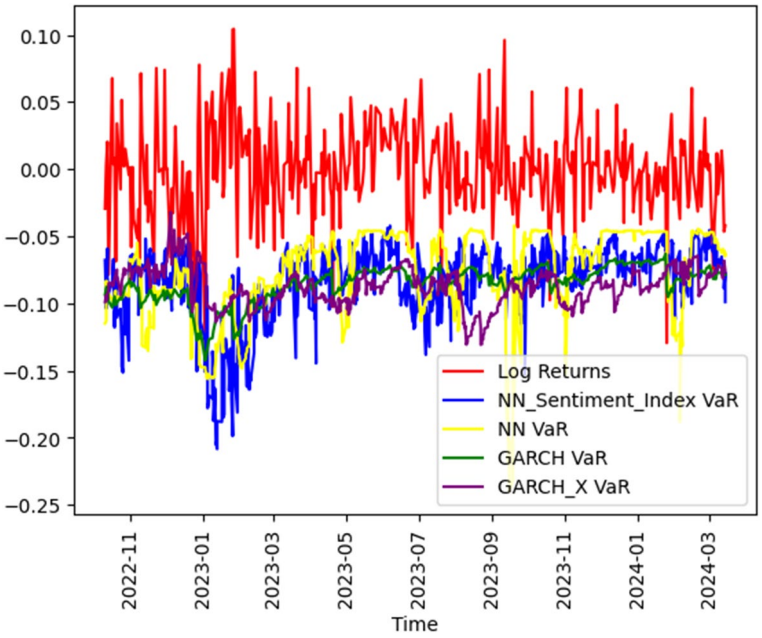

Fig. 5 Forecasts of the 2.5% Value-at-Risk for Tesla Inc. stock in the test dataset as predicted by multiple models

the different neural network models are substantially greater than those of the other models' forecasts. However, the instance of the *NN* model shows that the flexibility provided by the neural network alone is not sufficient. The additional information provided by the different sentiment indices also increases the accuracy of the VaR forecasts which ensures the immediate response to most sudden changes with appropriate absolute magnitudes. The $NN\_Sentiment\_Index$ and $GARCH\_X$ models establish lower VaR levels in high volatility periods. However, similar to the *NN* model, during less volatile periods, the $NN\_Sentiment\_Index$ model predicts higher VaR levels, which prevents the model from overestimating the VaR during these times, as is often the case with *GARCH* models (Lee & Su, 2012). This suggests that the adequate integration of sentiment data in different models contributes to a more conservative risk assessment in high-volatility periods, which prevents the underestimation of Value-at-Risk and the flexibility provided by the neural network ensures that the VaR level is not overestimated during less volatile periods. Based on the selected visual examples, it is evident that the $NN\_Sentiment\_Index$ model performs the best among all the models, because this model combines the indispensable complexity offered by the neural networks with the advantages of providing additional information on sentiment indices.

In addition to evaluating the visual results, we also conduct various statistical tests on the different VaR forecasts. The discrepancies from realized volatility for these stocks, the statistical backtests of the VaR, and the expected losses determined by various loss functions are summarized in Tables 10, 11, and 12. Table 10 reveals that the $NN\_Sentiment\_Index$ model consistently outperformed the other models in

Table 10 Metrics measuring the forecasting ability of volatility predicted by different models

	NN_Sentiment			NN			GARCH			GARCH_X		
	MSE	RMSPE	MA	MSE	RMSPE	MA	MSE	RMSPE	MA	MSE	RMSPE	MA
GOOG	4.504e−04	0.9768	1.966e−02	4.534e−04	0.9810	1.973e−02	4.547e−04	0.9786	1.973e−02	4.598e−04	0.9768	1.976e−02
IBM	1.663e−04	0.9674	1.174e−02	1.698e−04	0.9876	1.193e−02	1.685e−04	0.9797	1.186e−02	1.693e−04	0.9771	1.185e−02
TSLA	1.279e−03	0.9410	3.384e−02	1.297e−03	0.9602	3.431e−02	1.305e−03	0.9586	3.435e−02	1.310e−03	0.9524	3.343e−02

MSE denotes the Mean Square Error, RMSPE denotes the Root Mean Squared Prediction Error and MAE denotes the Mean Absolute Error. The different expressions of the error functions can be found in Formula 3

forecasting realized volatility for all randomly selected stocks. A pairwise comparison of the models revealed that the $NN\_Sentiment\_Index$ model performed better in most cases across all three metrics and outperformed in at least two metrics. For the randomly selected stocks under investigation, distinguishing among the remaining three models is not straightforward. Consequently, the comparison among these models is presented in the aggregated results section, which encompasses a larger sample size to provide a comprehensive analysis.

One of the most critical aspects of Value-at-Risk (VaR) backtesting is compliance with statistical tests. In the case of randomly selected stocks Table 11 shows that the GARCH model fails to meet any of the observed statistical tests at the 5% significance level, which confirms the general finding that GARCH models tend to underestimate or overestimate the actual Value-at-Risk (Lee & Su, 2012). In contrast, nearly all the other models satisfy the Kupiec test at the 5% significance level. However, only the VaR predictions generated by the $NN\_Sentiment\_Index$ model satisfy both the Christoffersen Duration and Joint tests. The observed results for the selected stocks are also reflected in the aggregated outcomes, demonstrating that only the $NN\_Sentiment\_Index$ model neither underestimates nor overestimates the Value-at-Risk, while the time intervals between violations exhibit no memory effect.

Finally, we assess the expected loss values estimated using different loss functions, as presented in Table 12. The table indicates that, in almost all cases, most of the different losses calculated with the VaR predictions from the *GARCH* model are the smallest. This is likely because although the *GARCH* model either underestimates or overestimates the VaR in certain periods, it generally tends to overestimate the VaR for the stocks examined during the whole period. The second smallest magnitude of the loss is achieved by the models enhanced with sentiment indices. While the traditional GARCH models appear to be the best in terms of minimizing expected losses, a broader perspective reveals a different conclusion. When considering which models achieve the lowest losses while also complying with various statistical tests, the $NN\_Sentiment\_Index$ model stands out as the best. For the examined stocks, the *GARCH* model does not satisfy any statistical tests, while the $NN\_Sentiment\_Index$ model not only adheres to the stringent requirements of statistical validation but also provides competitive loss values, making it the preferred choice for accurate and reliable Value-at-Risk forecasting.

6.3 Aggregated Results of Stocks in the S&P 500 Index

The analysis conducted on the randomly selected stocks is extended to include all stocks in the S&P 500 index. The aggregated results of this comprehensive analysis are presented in this section. First, we compare the different models pairwise to determine the proportion of stocks where each model more accurately predicts realized volatility relative to the others based on the specified metric. Second, we assess the proportion of stocks where one model realized less expected loss compared to the other model based on the specified metric. Finally, we evaluate the proportion of stocks for which one model more or less frequently rejected the null hypothesis than did the other model in the given statistical tests.

Table 11 Results of different statistical tests at the 5% significance level

	NN_Sentiment			NN			GARCH			GARCH_X		
	Kupiec	Christoffersen	Joint	Kupiec	Christoffersen	Joint	Kupiec	Christoffersen	Joint	Kupiec	Christoffersen	Joint
GOOG	FR H0	FR H0	FR H0	FR H0	R H0	R H0	R H0	R H0	R H0	FR H0	R H0	R H0
IBM	FR H0	FR H0	FR H0	FR H0	R H0	R H0	R H0	R H0	R H0	R H0	R H0	R H0
TSLA	FR H0	R H0	FR H0	FR H0	R H0	R H0	R H0	R H0	R H0	FR H0	R H0	R H0

We denote FR 0 to indicate that we failed to reject the null hypothesis of a given statistical test at the 5% significance level, while R 0 denotes that we rejected the null hypothesis of a given statistical test at the 5% significance level

The results are presented in Tables 13, 14, and 15, where the first two tables show the proportions of observed stocks determined by comparing how often the model in the given row outperforms the model in the corresponding column based on the specified metric. For instance in Table 13, the proportion of 69.57% demonstrates that the $NN\_Sentiment\ index$ model has a lower MSE than does to the *NN* model. This indicates that the $NN\_Sentiment\ index$ model outperforms the *NN* model in terms of the MSE for approximately 69.57% of the observed stocks evaluated. Naturally, if we examine the *NN* row, it becomes evident that in $1 - 69.57\% = 30.43\%$ of the cases, the *NN* model achieves a lower MSE value compared to the $NN\_Sentiment\_Index$ model. In contrast, the proportions presented in Table 15 should be interpreted differently. These proportions indicate how often the model in the given row performs better or worse than the model in the corresponding column. Hence, these values represent relative differences in proportions. For instance, in Table 15 the proportion of -6.38% at the intersection of the $NN\_Sentiment\ index$(row) and *NN* (column) for Kupiec means that in 6.38% of the stocks, the $NN\_Sentiment\ index$ model rejects the null hypothesis less frequently than does the *NN* model. A negative value indicates that the *NN* model outperforms the $NN\_Sentiment\ index$ model in terms of the Kupiec statistical test. The average magnitude of the volatility prediction errors of the different models compared to that of the GARCH model is presented in Fig. 6.

The $GARCH\_X$ model exhibits much better performance in terms of error metrics compared to the *GARCH* model in the case of the prediction of realized volatility. Figure 6 shows that the $GARCH\_X$ model exhibits significantly smaller average errors from realized volatility compared to the *GARCH* model based on all average metrics. Additionally, in nearly three-quarters of the examined stocks, the realized volatility is predicted more accurately by incorporating additional sentiment information compared to the *GARCH* model. This ratio is reversed when examining expected losses based on various metrics. In this case, the *GARCH* model predicts smaller losses in three-quarters of the instances compared to the $GARCH\_X$ model. However, according to the statistical tests, the $GARCH\_X$ model demonstrates less frequent underestimation or overestimation of the predicted VaR values. The $GARCH\_X$ model rejects the null hypothesis of Kupiec test less frequently than the *GARCH* model in approximately 15% of the stocks (75 stocks) and performs better in other statistical tests as well. Overall, by incorporating additional information derived from different sentiments, we can achieve better predictive capability for realized volatility in significantly more stocks and increase compliance with statistical tests, though this comes at the cost of potentially higher expected losses for certain stocks.

Comparing the *NN* model to the *GARCH* model, the GARCH(1,1) based *NN* model can also predict smaller deviations from realized volatility than the simple *GARCH* model in nearly half of the examined stocks, although the average error values of the *NN*'s predictions are greater than those of the *GARCH* model. This is because while the *NN* model provides better predictions in the majority of cases, the improvement in these instances is marginal. These marginal improvements are offset by the cases where the *GARCH* model performs better. Consequently, the average error values for the *NN* model are higher. One possible explanation for why the

Table 12 Metrics measuring the expected loss of Value-at-Risk predictions predicted by different models

	NN_Sentiment			NN			GARCH			GARCH_X		
	QL	FL	SL	QL	FL	SL	QL	FL	SL	QL	FL	SL
GOOG	1.960e−02	6.142e−02	−8.006e−03	2.521e−02	6.748e−02	−7.607e−03	8.394e−03	5.197e−02	−8.629e−03	2.521e−02	6.689e−02	−7.933e−03
IBM	7.026e−03	4.443e−02	−8.681e−03	1.967e−02	4.714e−02	−7.567e−03	5.618e−03	3.774e−02	−8.599e−03	8.429e−03	4.062e−02	−8.554e−03
TSLA	1.959e−02	1.090e−01	−5.600e−03	3.359e−02	1.107e−01	−5.488e−03	1.119e−02	9.555e−02	−6.095e−03	2.800e−02	1.137e−01	−4.987e−03

QL denotes the Quadratic Loss found in Formula 24, FR denotes the Firm Loss found in Formula 25, and SL denotes the Smooth Loss found in Formula 26

hybrid model does not perform better than the simple model is that, despite the use of multiple regularization techniques, the neural network model may have overfitted the remaining stocks. The *NN* model does not show significant improvement in reducing expected losses compared to the standard *GARCH* model, achieving lower expected losses in only approximately 10% of the examined stocks. Nevertheless, the results of the statistical tests improved significantly, for instance, the Kupiec test was satisfied more frequently in 35% of the examined cases (175 stocks), and Christoffersen's duration and Joint tests were also satisfied more frequently in one-quarter of the stocks. It is also observed that the *NN* model accepts the various statistical tests in significantly more instances than does the $GARCH\_X$ model. Overall, the neural network-based hybrid model demonstrates better volatility forecasting capabilities in many cases and satisfies statistical tests significantly more often than do the GARCH-based models; however, this comes at the cost of higher expected losses in many instances. The NN model's potential is evident, but additional enhancements are required to fully realize its capabilities.

The $NN\_Sentiment\_Index$ model combines all the advantages contained in the previously presented $GARCH\_X$ and *NN* models. While the neural network ensures the appropriate complexity of the model, the additional information derived from various sentiments provides the opportunity to identify unexpected changes. Consequently, the $NN\_Sentiment\_Index$ model can respond appropriately to both expected and unexpected changes not only in terms of timing but also in absolute magnitude, in contrast to the other examined models. The $NN\_Sentiment\_Index$ model volatility forecasting capability surpasses the other models' ability. When comparing pairwise with the other models, it achieves better results in approximately three-quarters of the examined stocks compared to the *NN* model and demonstrates superior predictive capability for approximately three-fifths of the examined stocks compared to the different GARCH models. Furthermore, the average error value relative to the *GARCH* model is notably significant, being more than twice as large as the improvement achieved by the $GARCH\_X$ model. In terms of expected losses, the model positions itself between the *NN* and $GARCH\_X$ models. It significantly predicts lower expected losses in more cases, based on various metrics, than does the *NN* model. However, compared to the $GARCH\_X$ model, it forecasts lower expected losses in only approximately one-third of the cases. Statistically, there is no evidence of a significant discrepancy in the proportion of stocks for which the *NN* model meets the criteria of the Kupiec test. However, substantial improvement is observed in the duration and joint statistical tests. The model satisfies the duration test in approximately an additional 20% of the examined stocks (almost 100 stocks) and the joint test in an additional 10%(more than 50 stocks) compared to the *NN* model. Ultimately, the $NN\_Sentiment\_Index$ model is the one that most accurately predicts volatility across the largest number of observed stocks among all the proposed models. Furthermore, the Value-at-Risk predicted by the $NN\_Sentiment\_Index$ model ensures the smallest expected loss for the majority of the observed stocks while satisfying the different statistical tests.

Table 13 The proportion of stocks for which the realized volatility is better predicted by the given model than by the other models based on the specified metric

	NN_Sentiment_Index	NN	GARCH	GARCH_X
NN_Sentiment_Index				
MSE		69.57%	63.19%	62.34%
RMSPE		79.36%	66.17%	54.47%
MAE		76.38%	65.53%	56.81%
NN				
MSE	30.43%		49.79%	43.19%
RMSPE	20.64%		37.66%	22.34%
MAE	23.62%		44.04%	27.87%
GARCH				
MSE	36.81%	50.21%		38.09%
RMSPE	33.83%	62.34%		24.04%
MAE	34.47%	55.96%		26.81%
GARCH_X				
MSE	37.66%	56.81%	61.91%	
RMSPE	45.53%	77.66%	75.96%	
MAE	43.19%	72.13%	73.19%	

Table 14 The proportion of stocks for which the loss is smaller than the given model compared to the other model based on the specified metric

	NN_Sentiment_Index	NN	GARCH	GARCH_X
NN_Sentiment_Index				
Quadratic Loss		88.30%	15.96%	34.26%
Firm Loss		60.00%	17.87%	30.21%
Smooth Loss		68.51%	23.83%	35.74%
NN				
Quadratic Loss	10.85%		7.45%	13.19%
Firm Loss	40.00%		14.04%	26.17%
Smooth Loss	31.49%		11.91%	24.68%
GARCH				
Quadratic Loss	80.00%	92.13%		69.79%
Firm Loss	82.13%	85.96%		78.30%
Smooth Loss	76.17%	88.09%		73.83%
GARCH_X				
Quadratic Loss	63.19%	86.17%	22.13%	
Firm Loss	69.79%	73.83%	21.70%	
Smooth Loss	64.26%	75.32%	26.17%	

7 Conclusion and Further Suggestions

In this paper, we propose incorporating the additional information provided by sentiments of news and social media into well-known conditional volatility forecasting models. The underlying idea is that sentiment indices can predict unexpected changes that cannot be anticipated based solely on information on previous returns. First, we explore the nonlinear functional relationship between various sentiment indices and

Table 15 The proportion of stocks for which the null hypothesis of the given test statistics was rejected more or less frequently compared to the other model

	NN_Sentiment_Index	NN	GARCH	GARCH_X
NN_Sentiment_Index				
Kupiec		−6.38%	28.72%	14.04%
Christoffersen		20.00%	44.04%	37.44%
Joint		11.06%	33.19%	27.87%
NN				
Kupiec	6.38%		35.11%	20.43%
Christoffersen	−20.00%		24.04%	17.45%
Joint	−11.06%		22.13%	16.81%
GARCH				
Kupiec	−28.72%	−35.11%		−14.68%
Christoffersen	−44.04%	−24.04%		−6.60%
Joint	−33.19%	−22.13%		−5.32%
GARCH_X				
Kupiec	−14.04%	−20.43%	14.68%	
Christoffersen	−37.45%	−17.45%	6.60%	
Joint	−27.87%	−16.81%	5.32%	

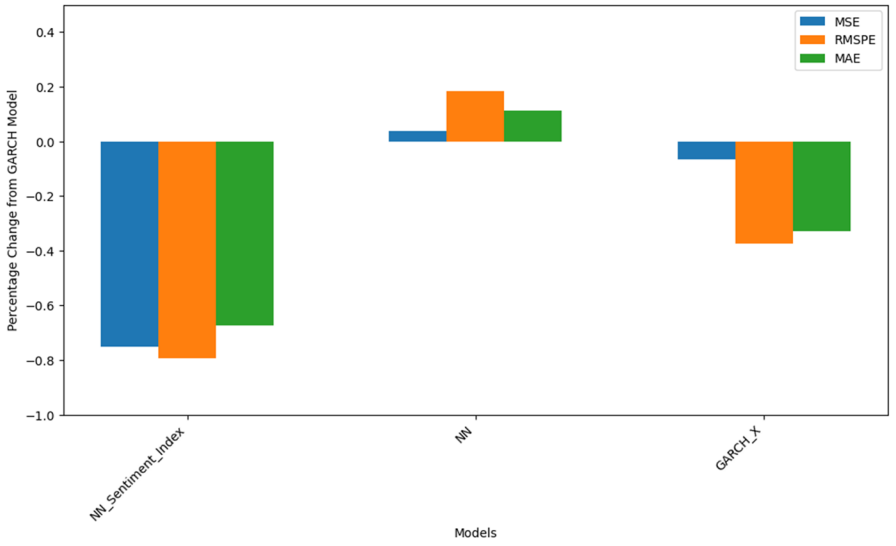

Fig. 6 The volatility prediction capability of the given models compared to that of the GARCH model

conditional volatility. Following this, we propose two models enhanced with sentiment indices to achieve more accurate conditional volatility predictions, and thereby more precise Value-at-Risk estimations.

We conducted an empirical analysis to compare different models. During the analysis, we estimate Value-at-Risk (VaR) values at a 2.5% significance level for all stocks listed in the S&P 500. The models are trained and backtested over a five-year period. Based on the predictions the sentiment augmented GARCH model is

more effective at predicting realized volatility across various metrics than the standard GARCH model. The additional information from sentiment indices reduces instances of underestimation or overestimation of the actual Value-at-Risk. GARCH prediction-based neural networks are capable of responding promptly to rapid returns changes. However, we demonstrate that these reactions are sometimes missing or often insufficiently large for accurate VaR estimation. Our proposed sentiment augmented neural network model combines the necessary complexity of neural networks and the ability to detect unexpected changes provided by the additional information of different sentiment indices. The model, therefore, is capable of responding not only promptly to unforeseen changes but also to changes of an appropriate magnitude.

Based on our results, the $NN\_Sentiment\_Index$ model most accurately predicts conditional volatility among all the models examined, including the standard GARCH, sentiment augmented GARCH and hybrid GARCH-LSTM approaches used as benchmarks. Furthermore, for most of the examined stocks, the Value-at-Risk estimated by the model is the one that realizes the smallest expected loss while satisfying the conventional statistical tests of Value-at-Risk. These features of the sentiment augmented neural network model demonstrate its superior performance and reliability in financial risk prediction compared to the popular benchmark models. Therefore, we recommend further development and integration of this model into risk management practices to enhance predictive accuracy and minimize potential financial losses.

During the training process for individual stocks, we refrained from performing hyperparameter optimization. We apply the same sufficiently complex architecture and hyperparameters in all cases to ensure that the estimated results from the models are easily comparable. Therefore, we believe that the results could be further improved if hyperparameter optimization is conducted separately for each stock. Unfortunately, we were only able to conduct the training and backtesting of the models over the past five years. However, if one has access to older sentiment data, it would also be possible to train the proposed models in a stressed environment, which we believe would result in even more conservative VaR predictions. In this study, we exclusively forecast 1-day ahead Value-at-Risk, but the architecture of the neural network with LSTM cells could enable longer-term forecasts, which, in our opinion, could achieve even greater improvement compared to conventional models. The models could further be expanded by incorporating information from other media sources, such as Facebook or Twitter. As suggested by Akbiyik et al. (2023), it might also be worthwhile to consider not only the content of the tweets but also the identity of the individual tweeting. Furthermore, the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method proposed by Zhang et al. (2024) could be applied before model training to achieve further performance improvements.

Acknowledgements The authors are grateful for the discussions with Péter Csóka, András Fülöp and Zsolt Darvas. We also extend our gratitude to the referees for their valuable suggestions.

Author Contributions Both authors contributed equally to the research.

Funding Open access funding provided by Corvinus University of Budapest. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability All data utilized in this study is available on the Bloomberg Terminal.

Code Availability The codes that were used in this study are available from the corresponding author upon request.

Declarations

Competing interests The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for Publication Not applicable.

Materials Availability Not applicable.

References

Akbiyik, M. E., Erkul, M., Kämpf, K., Vasiliauskaite, V., & Antulov-Fantulin, N. (2023). Ask "who", not "what": Bitcoin volatility forecasting with twitter data. *Proceedings of the sixteenth ACM international conference on web search and data mining* (pp. 688–696).

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.

Arslan, S. (2024). Bitcoin price prediction using sentiment analysis and empirical mode decomposition. *Computational Economics,* 1–22.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*(3), 203–228.

Banerjee, S., & Green, B. (2015). Signal or noise? Uncertainty and learning about whether other traders are informed. *Journal of Financial Economics, 117*(2), 398–423. https://doi.org/10.1016/j.jfineco.2015.05.003

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Bayer, S. (2019). *Var-backtesting*. https://github.com/BayerSe/VaR-Backtesting.git. GitHub.

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

Buczynski, M., & Chlebus, M. (2023). Garchnet: Value-at-risk forecasting with garch models based on neural networks. *Computational Economics,* 1–31.

Campbell, S. D. (2005). *A review of backtesting and backtesting procedures* (Finance and Economics Discussion Series No. 2005-21). Board of Governors of the Federal Reserve System (U.S.). Retrieved from https://EconPapers.repec.org/RePEc:fip:fedgfe:2005-21

Chen, X., & Hu, Y. (2022). Volatility forecasts of stock index futures in China and the us-a hybrid LSTM approach. *Plos One, 17*(7), e0271595.
Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record, 88*, 2–9.
Christensen, K., Siggaard, M., & Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics, 21*(5), 1680–1727.
Christoffersen, P., & Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics, 2*(1), 84–108.
Committee, B. (2013). Fundamental review of the trading book: A revised market risk framework. 11.04.2024, Retrieved 11.04.2024. https://www.bis.org/publ/bcbs265.pdff
Committee, B. (2019). Minimum capital requirements for market risk. 11.04.2024. Retrieved 11.04.2024. https://www.bis.org/bcbs/publ/d352.pdf
Cont, R. (2007). Volatility clustering in financial markets: Empirical facts and agent-based models. *Long memory in economics* (pp. 289–309). Springer.
Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance, 1*(2), 223.
de Vries, C. G., Samorodnitsky, G., Jorgensen, B. N., Mandira, S., & Danielsson, J. (2005). *Subadditivity re–examined: The case for value-at-risk* (Tech. Rep.) Financial Markets Group.
Degiannakis, S., Floros, C., & Livada, A. (2012). Evaluating value-at-risk models before and after the financial crisis of 2008: International evidence. *Managerial Finance, 38*(4), 436–452.
Deveikyte, J., Geman, H., Piccari, C., & Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence, 5*, Article 836809.
Donaldson, R., & Kamstra, M. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance, 4*(1), 17–46. https://doi.org/10.1016/S0927-5398(96)00011-4
Duffie, D., & Pan, J. (1997). An overview of value at risk. *The Journal of Derivatives, 4*(3), 7–49.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica,* 987–1007.
Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance, 48*(5), 1749–1778.
García-Medina, A., & Aguayo-Moreno, E. (2024). LSTM-GARCH hybrid model for the prediction of volatility in cryptocurrency portfolios. *Computational Economics, 63*(4), 1511–1542.
González-Rivera, G., Lee, T.- H., & Mishra, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting, 20*(4), 629–645.
Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and google trends. *Journal of Economic Behavior & Organization, 117*, 62–81.
Hamilton, J. D. (2020). *Time series analysis*. Princeton University Press.
Hu, Y., Ni, J., & Wen, L. (2020). A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction. *Physica A: Statistical Mechanics and its Applications, 557*, 124907. https://doi.org/10.1016/j.physa.2020.124907
Iqbal, N., Gul, F., & Mubarik, F. (2023). Investor sentiments and stock returns: A study on noise traders. *Journal of Positive School Psychology, 7*(1), 53–64.
Kakade, K., Jain, I., & Mishra, A. K. (2022). Value-at-risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach. *Resources Policy, 78*, 102903. https://doi.org/10.1016/j.resourpol.2022.102903
Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103*, 25–37. https://doi.org/10.1016/j.eswa.2018.03.002
Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *International conference on learning representations (ICLR)*. San Diego, CA, USA.
Kirman, A., & Teyssiere, G. (2002). Microeconomic models for long memory in the volatility of financial time series. *Studies in Nonlinear Dynamics & Econometrics,5*(4).
Kranefuss, E., & Johnson, D. K. (2021). Does twitter strengthen volatility forecasts? Evidence from the s &p 500, DJIA and twitter sentiment analysis. *Evidence from the S &P,500*.
Lee, C.- F., & Su, J.- B. (2012). Alternative statistical distributions for estimating value-at-risk: Theory and evidence. *Review of Quantitative Finance and Accounting,39*, 309–331.

Liu, Y. (2019). Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. *Expert Systems with Applications,132*, 99–109. https://doi.org/10.1016/j.eswa.2019.04.038. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417419302635

Lopez, J. A., et al. (1999). Methods for evaluating value-at-risk estimates. *Economic Review, 2*, 3–17.

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

Lu, X., Que, D., & Cao, G. (2016). Volatility forecast based on the hybrid artificial neural network and GARCH-type models. *Procedia Computer Science, 91*, 1044–1049.

Lux, T., & Marchesi, M. (1998). Volatility clustering in financial markets: A micro-simulation of interacting agents. *IFAC Proceedings Volumes,31*(16), 7–10. https://doi.org/10.1016/S1474-6670(17)40450-2. Retrieved from https://www.sciencedirect.com/science/article/pii/S1474667017404502. (IFAC Symposium on Computation in Economics, Finance and Engineering: Economic Systems, Cambridge, UK, 29 June - 1 July)

Morgan, N., & Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in neural information processing systems* (vol. 2).

Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R., & La Rocca, M. (2022). Deep learning for volatility forecasting in asset management. *Soft Computing, 26*(17), 8553–8574.

Rosoł, M., Młyńczak, M., & Cybulski, G. (2024). *Nonlincausality*. https://github.com/mrosol/Nonlincausality.git. GitHub.

Rosoł, M., Młyńczak, M., & Cybulski, G. (2022). Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. *Computer Methods and Programs in Biomedicine, 216*, Article 106669. https://doi.org/10.1016/j.cmpb.2022.106669

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Sarma, M., Thomas, S., & Shah, A. (2003). Selection of value-at-risk models. *Journal of Forecasting, 22*(4), 337–358.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673–2681.

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization,107*, 730–743. https://doi.org/10.1016/j.jebo.2014.06.004. (Empirical Behavioral Finance)

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

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

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437–1467.

Wang, Y., Andreeva, G., & Martin-Barragan, B. (2023). Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *International Review of Financial Analysis, 90*, Article 102914.

Zhang, Y., Peng, Y., & Song, Y. (2024). Realized volatility forecasting for stocks and futures indices with rolling CEEMDAN and machine learning models. *Computational Economics,* 1–54.

Authors and Affiliations

Dániel Léber[1] · Balázs Egyed[1]

✉ Dániel Léber
daniel.leber@stud.uni-corvinus.hu

Balázs Egyed
balazs.egyed@stud.uni-corvinus.hu

[1] Institute of Finance, Corvinus University of Budapest, Fővám Square 8, Budapest 1093, Budapest, Hungary