Journal of Economic Theory and Econometrics, Vol. 35, No. 1, Mar. 2024, 29–58

Forecasting Stock Market Volatility: A Sentiment Based Approach*

Gyujin Choi† Chang Sik Kim‡

Abstract This paper examines the impact of investor sentiment on stock market volatility using a natural language processing classification method applied to a large-scale dataset of social network data. We also apply numerous forecasting techniques not only including conventional linear models, but also different machine learning models and compare its results. Among various economic and sentiment features, we employ the least absolute shrinkage and selection operator (Lasso) for linear models and a tree-based nonlinear variable selection method to demonstrate the critical role of sentiment measures in market volatility. The results show that sentiment variables are identified to be one of the most important variables in relationship with stock market volatility and improve the future prediction of volatility when considered.

Keywords Forecasting market volatility, investor sentiment, machine learning, VIX Index.

JEL Classification C55, L50.

*We are grateful to the editor and two anonymous reviewers for their thoughtful comments and suggestions.

†Department of Economics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Republic of Korea 03063. Email: gyjin25@gmail.com

‡Corresponding Author. Department of Economics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Republic of Korea 03063. Email: skimcs@skku.edu

Received February 16, 2024, Revised March 14, 2024, Accepted March 16, 2024

1. INTRODUCTION

It is well known that VIX index (Chicago Board Options Exchange Volatility Index) is designed to measure the expectation of 30-day volatility implied by S&P500 prices of puts and calls. Considering that VIX index represents expected market fluctuations, it is crucial for informing market participants about how the market is reacting to future anticipated risks. Moreover, with the financial market becoming more sophisticated and various derivatives emerging with all kinds of underlying assets, VIX futures and options were launched in 2004 and 2006, providing efficient information on how participants in financial markets perceive. Therefore, deliberate predictions of the VIX index will provide valuable information to investors, alleviating information asymmetry among market participants.

The Efficient Market Hypothesis maintains that price reflects all available information, but numerous empirical findings suggest anomalies in many of financial markets. This paper mainly studies how investors' sentiment, collected from social network service and classified by natural language processing techniques, effects the volatility of stock market. Over 2.5 million observations of investor sentiment data have been collected from Twitter over the past 10 years and classified using the BERT algorithm proposed by Devlin *et al.* (2018). This paper modifies the heterogeneous autoregressive (HAR) model proposed by Corsi (2009), not only by using different covariates but also by adapting HAR into different machine learning algorithms. Along with the sentiment data, numerous economic covariates (Equity Market, Bond and Exchange Market, Liquidity, and Macroeconomic) are also considered regarding its impact on stock market volatility. Not only conventional linear regression-based methodologies, but various machine learning methodologies including Lasso based models, Random Forest, and XGBoost are considered to determine sentiment variable's impact on future volatility forecasting in both linear and nonlinear perspective. We also try to clarify whether the sentiment index extracted from social media is actually related and have a significant effect on stock market volatility with various models.

Our empirical findings show that economic and sentiment variables play a crucial role in future volatility forecasting, which coincides with the results from Audrino *et al.* (2020), where the authors implemented HAR based adaptive Lasso method for variable selection and prediction. They demonstrated that sentiment and attention variables have predictive power for the future volatility of individual stocks. However, this paper suggests that the relationship still holds when nonlinear relationship is allowed, demonstrating that sentiment covariates

are shown to be significant in Random Forest and XGBoost models. Moreover, forecasting stock market volatility with sentiment variables leads to lower MSEs in both linear models and nonlinear models reinforcing the argument that sentiment measures are significant predictors of future market volatility. However, the forecasting error in the short run decreases when utilizing sentiment variables in both linear and nonlinear models. The effects of sentiment covariates weaken and eventually disappear in market volatility forecasting beyond one or two weeks. While we cannot exclude the possibility that these results could potentially be driven by rational factors, such as unidentified state variables or market frictions, we believe that the predictive power of sentiment covariates is related to the activities of noise traders rather than influencing the behavior of institutional long-term investors.

The rest of this paper is as follows. Section 2 discusses the related previous literature and Section 3 describes the data set we use in this paper. In Section 4, we present various models considered in the paper and our main empirical results and forecasting performances. Section 5 provides a short conclusion.

2. LITERATURE REVIEW

The majority of studies have examined methodologies to accurately forecast market volatility, considering both specific securities based on realized volatility and implied volatility based on the VIX index. Christiansen *et al.* (2012) utilized economic variables to forecast realized volatility, verifying the predictive power of these variables. Fernandes *et al.* (2014) employed the HAR model to forecast the VIX index with numerous economic covariates, also attempting to capture the non-linear relationship using a neural network approach. Ballestra *et al.* (2019) forecast VIX futures using a feed-forward neural network, resulting in higher accuracy of predictions.

In a new direction, studies analyzing the behavior of the stock market using investor sentiment have been expanding, propelled by larger datasets and faster computing powers. Antweiler and Frank (2004) studied how Internet stock message boards was related to the stock market excess return, defining a disagreement measure among the messages to predict trading volumes. Cookson and Niessner (2019) gathered messages from Stocktwits, a social network platform where investors share their opinions on different financial securities. They classified over 1 million messages and derived the disagreement measure which Antweiler and Frank (2004) proposed, showing that disagreement among investors lead to a significant increase in abnormal trading volume. Seo and Kim

(2015) used HAR model to forecast realized volatility during high and low sentiment periods, using the sentiment index created by Baker and Wurgler (2006).

In addition, there have been innovative approaches applying machine learning methodologies in economic literature, particularly in forecasting problems. In case of forecasting market volatility utilizing machine learning, Hosker *et al.* (2019) argued that Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) has improved forecasting accuracy compared to other supervised learning methodologies including Lasso, Support Vector Regression (SVR), and Random Forests (RF). Vrontos *et al.* (2021) also showed that not only these numerous machine-learning techniques, including Elastic Net, Discriminant Analysis, Bayesian Models, K-Nearest Neighbors, and Forests with bagging and boosting improved accuracy of out-of-sample forecasting, but also the use of penalization terms plays a crucial role in the reduction of prediction errors. Concentrating more on neural network-based forecasting, Kim and Baek (2018) demonstrated neural network based HAR model generally performs better than the traditional HAR, but adding too many features might decrease its supremacy. Kim and Baek (2019) improves the results by Factor-Augmented HAR model using LSTM networks and showed augmentation of factors improves realized volatility forecasting with S&P, and numerous stock indices in Asia such as Nikkei, Hangseng, and KOSPI.

This paper mainly follows the framework of Audrino *et al.* (2020) with three major improvements. First, this paper investigates whether the investor sentiment has a significant influence not only on a specific firm level, but on general financial market volatility using VIX index as the target variable. In addition, this paper utilizes various machine learning techniques to verify the nonlinear effect of sentiment and economic variables, demonstrating that sentiment variables are selected in feature importance of Random Forest Models. Moreover, this paper considers various machine learning techniques, as well as linear adaptive-Lasso, for forecasting VIX, showing an improvement in forecast accuracy when sentiment variables are used. Remaining parts of this paper is organized as follows. Section 3 provides data and models, briefly introducing machine learning based techniques. Section 4 summarizes the results of variable selection and forecasting accuracy, and Section 5 concludes this paper.

3. DATA

Along with CBOE VIX index, this paper considers numerous economic and sentiment variables in order to verify whether investors' sentiment plays a crucial

role in forecasting and improves the accuracy of forecast itself.

The timespan of the daily data used in the paper spans from January 1, 2010, to August 31, 2022. However, considering the complexity of the sentiment data, specifically the collected twitter data which will be later described, this paper divides the timespan of the observations, resulting two different datasets. The first span of the dataset consists of daily observations from January 1, 2010 to December 31, 2018 and the second span includes observations from January 1, 2019 to August 31, 2022. Taking holidays and non-trading days into account, total 2,197 observations are used for the first dataset, and 841 observations are used for the second dataset. Considering the fact that some of raw data have different frequencies, interpolation and aggregation methodologies are applied. For forecasting, data until day t in Eastern time is used to predict the VIX index on day $t+1$.

Numerous economic variables are used in this research to predict the VIX index, many of which are also considered in previous literature, including Audrino *et al.* (2020) and Kim and Han (2022), among many others. It is widely known that macroeconomic and financial economic variables significantly are helpful to volatility forecasting VIX as well as realized volatility of the market. Following Audrino *et al.* (2020), this paper also categorizes macroeconomic variables into four different categories:

- Equity Market Variables: GSPC (S&P500) Returns, DJI (Dow Jones Industrial Average) Returns, MSCI (Morgan Stanley Capital International) Returns, Fama-French Factors (MKT-RF, SMB, HML), Short-term reversal.
- Bond & Exchange Market Variables: T-bill rate, Term-Spread, Credit Spread, FF-Deviation, log-return of spot exchange rate (EUR, GBP, JPY, CNY, CHF), and Dollar Index.
- Liquidity Variables: Turnover-Ratio of Dow Jones, Turnover-Ratio of change in Dow Jones, Turnover-Ratio of MSCI, Turnover-Ratio of change in MSCI, Turnover-Ratio of S&P, Turnover-Ratio of change in S&P.
- Macroeconomic Variables: Inflation Rate (Interpolated [1]), Industrial Production Growth (Interpolated), New Orders Growth for Durable Goods (Interpolated), Private Housing (Interpolated), Money Supply M1 (Interpolated), Consumer Sentiment of University of Michigan (Interpolated), CRB Spot Return, Capacity Utilization Level (Interpolated), and WTI Crude Oil price.

[1] Some of the macroeconomic time series are available only at a monthly frequency, and we use a linear interpolation of the monthly observations.

Along with different economic covariates, this paper also investigates the effect of investor sentiment on the VIX index. Audrino *et al.* (2020) considered both Twitter and StockTwits for daily sentiment measures, but we only use Twitter for gathering tweets since StockTwits limited their API access in 2021. We gathered a total of 2,064,414 raw tweets from Twitter itself using web scraping. Since our paper focuses on the overall financial market for VIX forecasting, broad keywords related to the entire financial market are used to select tweets. Tweets containing words 'S&P500', 'SP500', 'MSCI', 'Dow Jones', 'DJI', and '$SPY' are scrapped.

For each tweet, sentiment scores are computed by RoBERTa-based model (Loureiro *et al.* (2022)) from Hugging Face, which is trained for sentiment analysis using Twitter and trained on 124 million tweets. It classifies each tweet and assigns three labels: Negative, Neutral, and Positive, with each label having scores. For example, "What are you learning the methods that are being consumed by everyone. 95% of traders fail." is a tweet from April 1, 2016. The RoBERTa model computes its sentiment scores by assigning scores to each label, specifically, it assigned 0.727 to negative, 0.256 to neutral, and 0.017 to positive. For computing overall sentiment score for each tweet, scores computed by RoBERTa model is averaged by multiplying -1 to the negative score, 0 to the neutral score, and 1 to the positive score, making one sentiment score for each tweet. The RoBERTa model utilizes transformers for its computation, a deep learning model based on a neural network. However, unlike traditional RNN and LSTM based machine translation models which process sentence word by word, RoBERTa transformers are well known to have better performances, with whole sentence being processed and not suffering from long dependency problems.

Since Twitter users have increased significantly with the accessibility of mobile applications for social network services, it is natural that the number of extracted tweets increases over time. However, as twitter users increase and more tweets are posted, and naturally noisier and less informative tweets are also significantly increased. Typical examples are tweets on politics, AI-generated advertisements, cryptocurrency advertisements, and tweets on influencers.

Table 1 illustrates some of extracted tweets that are not directly related to stock market mentioning only political matters. It is evident that the content of these tweets does not reflect any beliefs or thoughts on the financial market. Instead, the users who posted these tweets added 'SPY' or 'SP500' so that their tweets can be seen by more people.

Table 2 illustrates the total number of extracted tweets and the number of

Date	Score	Content
2013.07.03	-0.9371	"I told you all. @BarackObama is a OOO Muslim. He is the ultimate OOO pos!!! $SPY $GLD"
2016.02.26	-0.9239	"Get that orange colored OOO off the screen. Trump sucks. $SPY"
2016.03.02	-0.9005	"This presidency is a sham and we are being pushed towards nationalism while zirp intends to enslave us #Trump #Sanders #Clinton $SPY"
2019.03.19	0.9553	"Really appreciating the live podcast from yesterday @peterschiff. It looks as if I'm only 40% way through, so more to hear but solid job. $SPY"
2019.09.29	0.9693	"Don't miss these crypto holders! Buy and Hold! $SPYthis is real gem! @satopay1 telegram @mdpienaar"
2021.12.21	0.0429	"[scan results – 15m]: #ftx 5 bullish trend (#perp futures)$mkr: 2 $alice top 5 bullish trend (#USD Stocks) follow @dyorcryptobot for more contents"
2022.01.28	0.0441	"Did you catch the flush on $SPY? .37 in 1.12 in minutes after my puts alert in the rooms. You need to be in the rooms for answer @stockpastor"

Table 1: EXAMPLE OF EXTRACTED TWEETS WITH UNRELATED CONTENTS.

tweets related to different politicians. After discarding the early periods, the total number of tweets remained around 100,000 to 150,000 until 2019, sharply increasing after 2020. Moreover, the number of tweets on politicians has increased significantly around 2018-2019, inducing notable negative sentiment scores compared to the mean score of total tweets. In addition, automatic twitter bots, advertising tweets of cryptocurrency trading platform, irrelevant political tweets after the midterm elections, and numerous other spamming tweets also have been increased.

We filtered out all irrelevant tweets, including advertisements, political content, influencer-related posts, and unrelated conversations. Specifically, we excluded tweets containing specific political keywords (Trump, Biden, Hillary, Obama, Republican, Democratic, Musk, Russia, Ukraine), tweets with cryp-

Year	Before	After	Trump	Obama	Hilary	Biden	Musk
2010	24,342	24,342	3	60	0	0	3
2011	66,098	66,098	24	348	2	5	8
2012	89,221	89,221	26	687	1	12	21
2013	113,605	113,605	27	522	0	9	33
2014	113,995	113,995	32	309	0	2	50
2015	153,414	153,414	100	141	37	5	30
2016	141,726	141,726	1,551	222	345	3	58
2017	90,834	90,834	2,838	190	26	0	47
2018	135,679	135,679	5,147	338	57	4	121
2019	113,965	51,243	4,973	76	19	41	118
2020	261,936	110,938	9,517	255	17	1,816	499
2021	284,028	110,135	824	32	7	2,850	566
2022	441,208	181,830	714	68	18	4,053	811

Table 2: NUMBER OF TWEETS FOR SPECIFIC KEYWORD.

tocurrency related terms (Crypto, Bitcoin, BTC, Ethereum, Doge, XRP), tweets with hyperlinks, retweeted content, and tweets that repeatedly share the same content. Additionally, for the second dataset with a timespan from January 2019 to August 2022, considering the steep increase in the number of noise tweets after 2018, tweets with sentiment scores between -0.3 and 0.3, and tweets with sentiment scores above and below 0.95 and -0.95, have also been filtered out to eliminate additional uncaptured cryptocurrency platform advertisements, automatic twitter bots, and unrelated chats.[2]

Figure 1 presents average daily sentiment scores that can be computed by averaging out all individual sentiment scores. It is evident that the variance of daily sentiment scores is high around 2010-2012 and gradually decreases over time, as shown in Figure 1. This can be attributed to the significantly fewer tweets on the stock and financial market in the early periods, with 24,342 tweets in 2010, 66,098 tweets in 2011, compared to 135,679 tweets in 2018. On the other hand, the high fluctuation in 2018 mirrors the stock market crash in February 2018, when the Dow dropped more than 12% in two weeks. It is also notable that the overall mean of the sentiment score is over 0, precisely 0.1014, indicating that people generally hold a positive belief in the financial market, consistent with

[2]If we split the second sample as of Jan. 2020, we obtain almost the same empirical results as those presented in Section 4. Therefore, the splitting point for the second sample is not important as long as it falls within the range of 2019-2020.

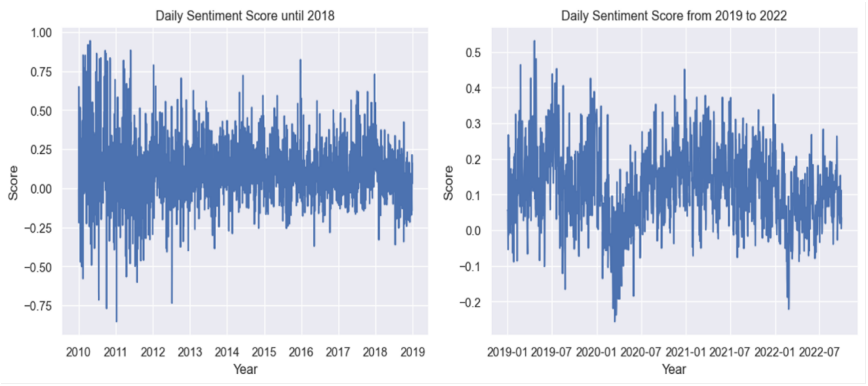

Figure 1: DAILY SENTIMENT SCORE FROM 2010 TO 2022.

the findings of Cookson and Niessner (2019). Moreover, a significant decline in sentiment scores is conspicuous during January 2020, primarily attributed to the impact of the COVID-19 pandemic.

Along with the sentiment score computed from Twitter, this paper also investigates the effects of different sentiment features, specifically Google search volume data. Google search volume data is obtained from Google Trends, which provides relative numbers of daily search queries within a 269-day period. We

Figure 2: GOOGLE SEARCH VOLUME FOR KEYWORDS.

adopt the method used by Audrino *et al.* (2020) to normalize the number of Google searches from the beginning of 2010 to the end of 2018, setting the highest day to 100 and the lowest day to 0. We consider the Google search volume for the following words: 'Stock Market', 'Financial Market', 'VIX', 'S&P', 'MSCI', 'Dow Jones Industrial Average'. While search volumes for 'Financial Market' and 'MSCI' exhibit considerable noise, other covariates tend to capture the movements of the financial market itself. All show extremely high search volumes in February 2018, as illustrated in Figure 2.

4. MODEL AND EMPIRICAL RESULTS

This section explores the significance of economic and sentiment covariates in volatility forecasting, employing both linear and nonlinear models. In the case of linear models, Lasso-based models are proposed with variable selection, and their forecasting power is compared to heterogeneous autoregression (HAR) models with economic and sentiment covariates. For nonlinear models, we employ tree-based machine learning models to compare HAR with sentiment and economic covariates. In particular, ensemble models using bagging (Random Forest) and boosting (XGBoost) are considered, with computation of variable importance and forecasting power. [3]

The benchmark HAR model introduced by Corsi (2009), is widely known as having high accuracy in predicting not only realized volatility, but also VIX. Numerous literatures already applied HAR models in forecasting VIX including Fernandes *et al.* (2014), Ballestra *et al.* (2019), Kim and Han (2022) among many others. The benchmark HAR model is given by

$$\begin{aligned} Model\ 1 : \log VIX(D)_{t+1} = \beta_0 + \beta_1 \log VIX(D)_t \\ + \beta_2 \log VIX(W)_t + \beta_3 \log VIX(M)_t + \varepsilon_{t+1}, \end{aligned} \tag{1}$$

where D represent daily, W and M represents weekly and monthly averages of daily log of VIX. Hereafter, Model 1 indicates the benchmark model with basic HAR covariates without any variable selections applied. We extend the basic HAR Model 1 by including equity market, bond market, exchange market, liquidity, and macroeconomic variables that are explained in the previous section. It can be written as

[3] As one referee suggested, incorporating the performances of neural Network-based machine learning models including LSTM would be valuable, but we leave those projects to our future research.

$$\begin{aligned} Model\ 2 : \log VIX(D)_{t+1} = \beta_0 + (\log VIX_t)'\beta_{vix} \\ + M_t'\gamma_{eco} + ... + M_{t-4}'\gamma_{eco} + \varepsilon_{t+1}, \end{aligned} \tag{2}$$

where lagged economics covariates are also considered up to $t-4$. Here, VIX_t denotes vector of $VIX(D)$, $VIX(W)$, $VIX(M)$ in the (1). We also extend Model 2 by adding Twitter sentiment scores and google search volumes for 6 keywords given in Section 3 as

$$\begin{aligned} Model\ 3 : \log VIX(D)_{t+1} = \beta_0 + (\log VIX_t)'\beta_{vix} \\ + M_t'\gamma_{eco} + ... + M_{t-4}'\gamma_{eco} \\ + S_t'\theta_{sent} + ... + S_{t-4}'\theta_{sent} + \varepsilon_{t+1}. \end{aligned} \tag{3}$$

Note that Model 1 is a basic HAR model with neither economic nor sentiment covariates, Model 2 includes economic covariates and their lagged terms only, and Model 3 incorporates all economic and sentiment covariates along with their lagged terms. After estimating Models 1, 2, and 3 using different machine learning methodologies, rolling forecasts for 1 step, 5 steps, 10 steps, and 22 steps ahead are generated to compare the effects of sentiment variables, following the approach in Audrino *et al.* (2020).

4.1. TRAINING AND TESTING SETS

In the context of training and testing sets, from the total of 2,197 observations in the first dataset (including observations until December 2018), the first 2,077 observations are utilized for training, while 120 observations are reserved for testing through one-day ahead direct forecasting using a rolling-window method. Considering that normalization or standardization is essential for Lasso-based linear models to match the scale of the penalty term. We employ min-max normalization and assess whether the results significantly differ from other standardization methods. To facilitate the interpretation and comparison of results, inverse scaling is implemented for forecasting and MSE computation.

Two different types of cross-validation methods are implemented for the training set. This paper employs the original k-fold cross-validation with 5 folds. In this method, one fold is randomly chosen as the validation set, and the remaining four folds constitute the training set for adjusting hyperparameters. However, given the time-dependent structure inherent in the series, there is a possibility that some splits might disrupt the dependent structure, leading to unrealistic values for hyperparameters. Thus, we also implement the time-series split from the

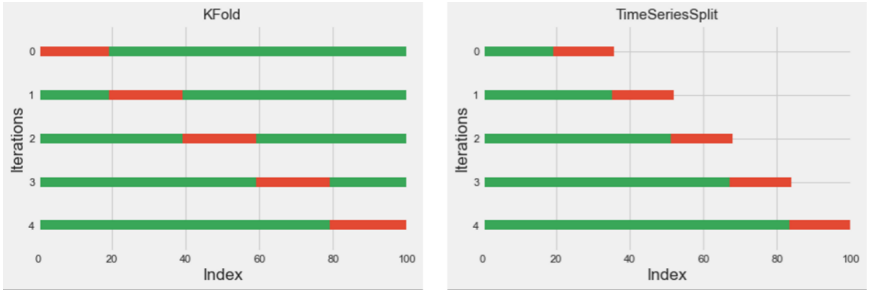

Figure 3: GOOGLE SEARCH VOLUME FOR KEYWORDS.

Scikit-learn library and consecutively verify the training and validation sets. For example, as illustrated in Figure 3, if the 1st, 2nd, and 3rd block are used for training, then the 4th block is automatically used for validation without compromising the dependent structure itself.

4.2. LINEAR MODELS WITH LASSO SELECTIONS

This paper first considers linear models, specifically Lasso models for variable selection and forecasting. Tibshirani (1996) introduces the regression shrinkage and variable selection by Lasso with L^1 regularization, only selecting features that have significant impact on the dependent variable by incorporating a constant penalty term in the original cost function of OLS. The hyperparameter λ is determined through cross-validation, employing both k-fold and time-series split methods, with values ranging from 0.001 to 2 in increments of 0.1, resulting in a total of 20 different λ. Through grid search, in both k-fold and time-series cross-validation, λ was determined to be 0.001. Three HAR features (daily, weekly, and monthly) are excluded for variable selection following Audrino *et al.* (2020). The Lasso estimator is given by

$$\hat{\beta}_L = \text{argmin}_\beta (y - X\beta)'(y - X\beta) + \lambda \|\beta\|_1 \tag{4}$$

and as mentioned earlier, note that X is a matrix of all covariates including economic and sentiment covariates, with lagged terms. Figure 4 demonstrates which variables are selected from LASSO selection method for all of 4 different forecasting horizons.

The red graphs in Figure 4 represent the selected sentiment variables, which include the Twitter sentiment score, Google search volumes, and their lagged

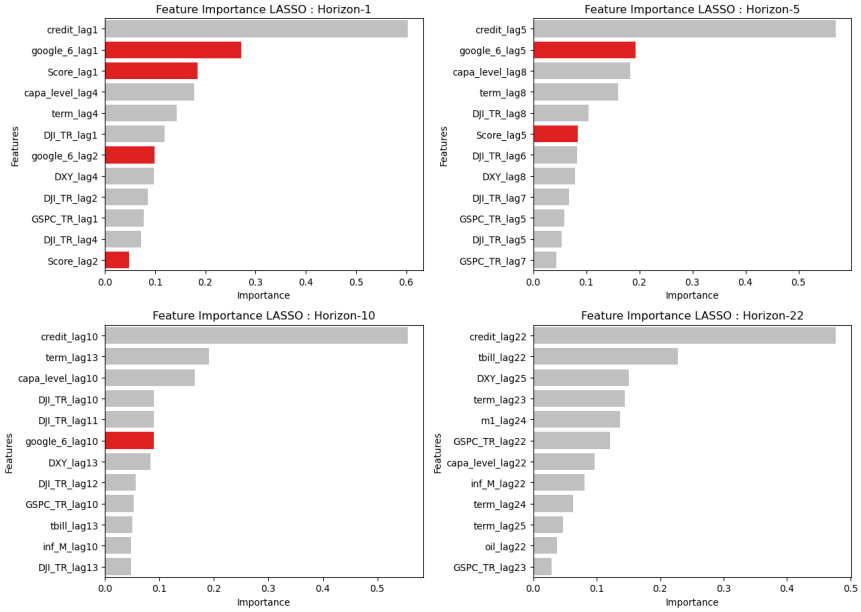

Figure 4: LASSO VARIABLE SELECTION RESULTS. The detailed descriptions of selected variables are given in Appendix, and lag# indicates the order of lags of corresponding variables.

terms. As the forecast horizon increases, the selected numbers of sentiment-related variables are decreasing. Focusing on 1-day ahead forecasting, the Google search volume for 'Dow Jones Industrial Average' and Twitter sentiment score are demonstrated to have a significant impact on the next day's volatility.

We also use adaptive Lasso selection by Zou (2006) for the variable selection. It is well known for giving heterogeneous penalty terms for each covariate, avoiding overfitting but also implementing shrinkage effect as Lasso. Regularization hyperparameter λ is also determined by two cross validation methodologies as in Lasso, and values ranging from 0.001 to 2 with 0.1 steps. Note that the adaptive weight term which performs the role of giving different penalty terms for each covariate, is used by the OLS values. The adaptive Lasso estimator is given by

$$\hat{\beta}_{AL} = \text{argmin}_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{n} \frac{|\beta_i|}{|\hat{\beta}_i|}. \tag{5}$$

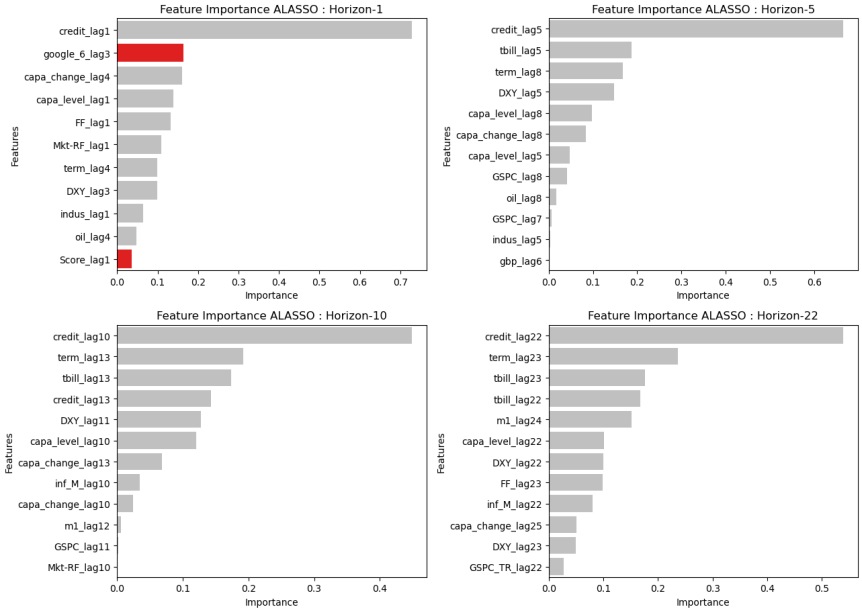

Figure 5: ADAPTIVE LASSO VARIABLE SELECTION RESULTS.

Figure 5 displays the outcomes of adaptive Lasso selection for various forecast horizons. It is noteworthy that fewer variables are selected compared to Lasso selection across all horizons. However, sentiment variables continue to be selected in the 1-day ahead forecast. Based on the selection outcomes of adaptive Lasso, it becomes evident that the sentiment effect plays a crucial role, particularly in the 1-day ahead forecast. Moreover, its impact sharply diminishes as forecast horizons increase. This trend aligns with other nonlinear forecasting results, which will be presented in subsequent sections.[4]

4.3. FEATURE IMPORTANCE WITH RANDOM FORESTS

This paper also explores the impact of sentiment and economic variables on volatility using nonlinear tree-based models. Random Forests, introduced by

[4]We also tried another linear variable selection method, namely, the group Lasso selection proposed by Yuan and Lin (2006)Given that the covariates investigated in this paper fall into 5 distinct groups, such as economic variables grouped by equity market, bond and exchange market, liquidity market, and macroeconomic variables, with the sentiment group comprising the Twitter sentiment score and Google search volumes. The results of group Lasso selection indicate that the sentiment group is consistently retained, even under group-penalization, across all forecast horizons. Here, we skipped the detailed results of group Lasso selection to save space.

Breiman (2001), is a well-known and popular machine learning technique that assembles numerous decision trees through bagging. Bagging is an algorithm that uses bootstrapping and aggregating individual decision trees to produce stable, non-overfitting results. From the original training set, m variables are randomly selected from the original training set, where $m < p$. It randomly chooses X_is, or features and samples uniformly with replacement to create new, or pseudo training sets from the original one. As sampling is conducted with replacement, some observations may be repeated, leading to a mixed sequence of original observations.

However, many papers implementing random forests overlook the fact that IID bootstrapping and the creation of pseudo training sets may disrupt the time-dependent structure in the given economic time series. In other words, bootstraps creating by uniform and independent random sampling may not fully represent the dependent structure of original training sets. In this paper, we attempt to implement the stationary bootstrap method introduced by Politis and Romano (1994) as a substitute for the original IID bootstrap method in generating random samples.

As explained in Section 4.1, we use both the k-fold and time-series split cross-validation methods for hyperparameter tuning. The number of estimators (number of trees) is set from 100 to, and the max depth (number of splits for each tree) is set from 4 to 14. With a total of 30 different parameters for grid search, hyperparameters of IID bootstrap-based random forest and stationary bootstrap-based random forest are separately determined for forecasting and computing variable importance.

In fact, we find that K-fold and time-series split cross-validation results turned out to be very similar, implying that the k-fold or time-series CV method shares a similar selection of best models. Furthermore, both stationary and IID bootstrap methods showed no significant difference, not only in variable selection but also in forecasting. In sum, considering time-dependent structure may not be an important in our volatility forecasting model with sentiment covariates.

The results of feature importance, computed using out-of-bag samples during the bagging procedure, are presented in Figure 6. Similar to the results of linear models with Lasso and adaptive Lasso, sentiment variables are among the top 12 in variable importance for short forecast horizons, but as the forecast horizon increases, sentiment variables have negligible impact on market volatility compared to the conventional economic variables. Both linear and tree-based models consistently show that sentiment covariates have a notable impact on forecasting future volatility, particularly in the 1-5-days ahead horizon. How-

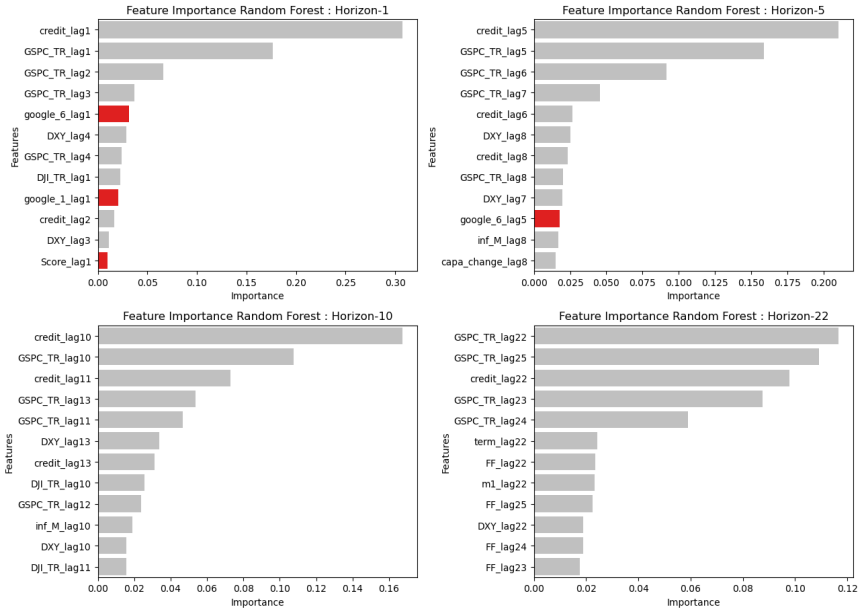

Figure 6: RANDOM FOREST VARIABLE IMPORTANCE RESULTS.

ever, this forecasting power diminishes as the horizons extend. Considering the heterogeneities among investors, the results confirm that sentiment variables are more representative of the behavior of individual investors and less indicative of the behavior of institutional investors, whose investment portfolios place more emphasis on long-term trading horizons.

4.4. OUT-OF-SAMPLE FORECASTING RESULTS

As shown in the previous sections, linear and nonlinear variable selection methodologies confirm that sentiment variables play a crucial role in volatility forecasting, specifically in short term forecasting. This section concentrates on out-of-sample forecasting to show whether the selection results actually increase the performance of forecasting using different forecasting methodologies.

Note that along with bagging algorithms, this paper also investigates the effects of sentiment variables using gradient boosting algorithms, by implementing XGBoost(XGB) methodology introduced by Chen and Guestrin (2016) Gradient boosting is a widely used machine learning algorithm belonging to the ensemble learning family. It initiates with a weak model and progressively builds a stronger model. Specifically, it begins with an initial decision tree model, where

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	.006	.059	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.006	.059	.006	.059	.006	.060	.006	.060
Model 3 LS	**.005**	**.055**	**.005**	**.056**	**.006**	**.057**	**.005**	**.054**
Model 1 RF	.007	.064	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	.007	.063	.007	.064	.007	.065	.007	.066
Model 3 RF	**.005**	**.056**	**.005**	**.058**	**.006**	**.062**	**.005**	**.058**
Model 1 XGB	.007	.067	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.007	.063	.008	.064	.008	.066	.008	.066
Model 3 XGB	**.007**	**.062**	**.006**	**.061**	**.006**	**.063**	**.007**	**.065**

Table 3: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 1. LS, RF, XGB indicate linear least squares, random forest, and XGBoost, respectively.

the second model aims to fit the residuals of the initial leading tree using the given features for boosting. This process is repeated until it reaches a point where training errors does not decrease anymore. Remembering that the gradient of L^2 loss function is the fitted value minus the observed true value, the algorithm is named Gradient Descent as it moves in the opposite direction of the gradient itself. XGB is an optimized implementation of the gradient boosting algorithm, renowned for its high performance and scalability.

In the following, models with different sets of covariates based on various variable selections will be utilized with linear least squares, Random Forest, and XGB methods, demonstrating whether sentiment variables can enhance forecasting performances. In fact, Model 1,2,3 in (1)-(3) presented in Section 4 with are all linear models. For expositional simplicity, we refer Model 1, 2, 3 for Random Forest and XGB as models with the same sets of features as in the linear Model 1,2,3. Hereafter, we use Model 1,2,3 notations in Random Forest and XGB as well. The following Table 3-6present 1,5,10,22-day ahead forecasting results

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	.030	.131	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.029	**.125**	.030	.128	**.032**	.133	.036	.135
Model 3 LS	**.028**	.126	**.029**	**.127**	.032	**.131**	**.036**	**.134**
Model 1 RF	.035	.144	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	.024	.119	.025	.118	**.021**	**.109**	.021	.110
Model 3 RF	**.023**	**.117**	**.024**	**.118**	.023	.114	**.020**	**.110**
Model 1 XGB	.037	.146	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.020	.109	**.013**	**.089**	.013	.091	**.018**	**.102**
Model 3 XGB	**.020**	**.108**	.014	.093	**.012**	**.085**	.018	.106

Table 4: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 5.

for the dataset, from January 1, 2010 to December 31, 2018, total with 2,197 observations.

The columns of Table 3 indicate the variable selection methods such as None, Lasso, Adaptive Lasso, and Random Forest. The first column (None) implies that no variable selection methodologies are implemented and uses all 168 variables to generate 1-day ahead forecast. In the Random Forest selection method, we choose the top 12 variables based on feature importance for the next day forecast. The rows of Table 3 indicate Model 1, Model 2, and Model 3 with different estimation procedures, i.e., least squares, Random Forest, and XGBoost as explained before. Model 1 includes the benchmark HAR covariates proposed by Corsi (2009). Model 2 incorporates 168 additional economic variables on top of those in Model 1. Lastly, Model 3 introduces extra sentiment covariates in addition to those in Model 2.

As shown in Table 3, the bold font represents the lowest MSE (Mean Squared Error) and MAE (Mean Absolute Error) in different forecasting methodologies

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.041**	.157	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.042	**.153**	**.038**	.154	**.037**	**.148**	**.053**	**.173**
Model 3 LS	.046	.160	.038	**.154**	.043	.166	.053	.175
Model 1 RF	.040	.165	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	**.025**	**.121**	**.021**	**.112**	.022	.112	**.022**	**.114**
Model 3 RF	.026	.124	.022	.115	**.021**	**.109**	.022	.116
Model 1 XGB	.040	.160	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.020	.107	**.016**	.103	.016	.098	**.016**	**.096**
Model 3 XGB	**.014**	**.097**	.017	**.101**	**.013**	**.087**	.018	.104

Table 5: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 10.

with various variable selections. It is evident that Model 3, enriched with sentiment covariates, produces the lowest MSE and MAE compared to Models 1 and 2 in 1-day ahead forecasting across both linear and tree-based models. This result shows that not just in variable selections, but also in out-of-sample forecast, sentiment variables play a crucial role improving forecasting errors of VIX for 1-day ahead forecast.

Tables 4-6 present the out-of-sample forecasting results for forecast horizons of 5, 10, and 22. We use direct forecasting rather than iterative method, which implies that no forecasting values are used for generating next step forecasts. In contrast to Table 3, where Model 3 consistently demonstrated superior forecasting performance regardless of the selection method and forecast methods, the dominance of Model 3 diminishes with increasing forecast horizons. The model with economic covariates usually shares similar MSE and MAE with the model with economic and sentiment covariates, which coincides with the results of variable selections in the previous sections. Sentiment variables' impact to

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.060**	**.189**	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.063	.205	**.060**	**.197**	**.060**	**.193**	**.050**	**.174**
Model 3 LS	.069	.218	.060	.197	.062	.194	.050	.175
Model 1 RF	.059	.202	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	**.020**	**.108**	**.015**	**.095**	**.012**	**.082**	**.016**	**.098**
Model 3 RF	.020	.109	.015	.095	.012	.084	.016	.099
Model 1 XGB	.060	.202	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	**.017**	**.093**	**.013**	**.089**	**.011**	**.077**	**.017**	**.096**
Model 3 XGB	.019	.101	.013	.089	.012	.086	.018	.097

Table 6: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 22.

future volatility was significant only in short forecasting horizon in the variable selections as well. Our findings clearly indicate a decrease in the forecasting power of sentiment variables as the prediction horizon extends. On average, the sentiment score computed from Twitter, Google search volume data improve volatility predictions slightly at the 1-5-days horizon, but cease to do so for prediction horizons over 10 trading days except XGB forecasting. XGB forecasting with sentiment covariates continues to outperform, even in 10-day ahead volatility forecasting. However, its forecasting efficacy diminishes as the prediction horizon extends beyond two weeks. In contrast, economic and financial variables exhibit predictive power for future volatility up to the 2-4 weeks horizon, as depicted in Tables 5 and 6. This differs from the findings of Audrino *et al.* (2020). Audrino *et al.* (2020) have found that, on average, economic variables have predictive power up to the two-week horizon, and sentiment covariates are able to increase the predictive accuracy for the one- and two-day-ahead predictions. However, in market volatility forecasting especially in nonlinear models,

the predictive power of sentiment and economic covariates lasts longer up to one- and four-week ahead forecasting. This is partly because linear HAR model cannot capture the complex nonlinearity between sentiment variables and market volatility.

4.5. THE EFFECTS OF SENTIMENT VARIABLE FOR A RECENT TIME PERIOD

Considering the impurity of Twitter sentiment, we have divided the entire dataset into two timespans. The second part of the dataset comprises observations from January 1, 2019, to August 31, 2022. As previously discussed in Section 3, the Twitter data are significantly affected by numerous extracted tweets containing political comments. These tweets often include 'SPY' or 'SP500' to attract a larger audience. Therefore, we rather use more thorough filters to remove those irrelevant tweets with advertisements, politics, influencers, and non-

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.004**	**.050**	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.010	.082	.005	.057	.005	.056	.005	.056
Model 3 LS	.009	.078	**.004**	**.049**	**.004**	**.050**	**.004**	**.048**
Model 1 RF	.006	.058	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	.005	.055	.005	.056	.005	.055	.005	.057
Model 3 RF	**.004**	**.049**	**.003**	**.044**	**.003**	**.045**	**.003**	**.046**
Model 1 XGB	.007	.064	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.007	.065	.007	.065	.006	.060	.005	.059
Model 3 XGB	**.004**	**.053**	**.004**	**.050**	**.004**	**.051**	**.004**	**.050**

Table 7: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 1.

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.018**	**.101**	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.036	.149	**.024**	**.128**	.023	.121	**.020**	.115
Model 3 LS	.037	.143	.026	.128	**.021**	**.115**	.021	**.108**
Model 1 RF	.021	.114	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	**.011**	**.084**	**.012**	**.086**	**.010**	**.079**	**.008**	**.073**
Model 3 RF	.011	.085	.012	.087	.010	.080	.009	.073
Model 1 XGB	.023	.120	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.013	.084	.012	.084	**.008**	.070	**.006**	**.062**
Model 3 XGB	**.011**	**.081**	**.011**	**.084**	.009	**.070**	.007	.065

Table 8: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 5.

related chats, tweets containing specific political keywords and crypto-currency keywords, etc.

Following the filtering processes, we employ the same methodologies and models to evaluate forecasting performances in future volatility. The COVID-19 period, spanning from February 1, 2020, to April 30, 2020, has been excluded from this dataset to avoid additional noise to the data. The total of 841 observations are utilized for the training with 120 observations reserved for the testing one-day ahead direct forecast.

Table 7 presents forecast results for the second part of the dataset, highlighting the overall dominance of Model 1 over Model 3 with sentiment covariates, except in the case of the linear model without variable selections. However, it is evident that, despite the prominence of Model 1, the inclusion of sentiment variables consistently enhances the forecasting performance for future volatility, particularly in relatively short horizons. Notably, after undergoing the relevant filtering processes, the improvement in forecasting power becomes even more

Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.028**	**.132**	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.045	.172	**.037**	**.158**	.039	.161	**.035**	**.161**
Model 3 LS	.056	.189	.040	.162	**.035**	**.155**	.040	.173
Model 1 RF	.029	.139	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	.017	.101	.016	.098	**.012**	.087	.012	**.084**
Model 3 RF	**.016**	**.100**	**.014**	**.092**	.012	**.086**	**.012**	.084
Model 1 XGB	.032	.143	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	.016	.099	.018	.099	.013	**.082**	.012	**.078**
Model 3 XGB	**.015**	**.096**	**.017**	**.093**	**.011**	.083	**.010**	.079

Table 9: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 10.

pronounced when compared to the results observed in the first part of the dataset, covering the period from 2010 to 2018.

Tables 8-10 present the forecasting results for 5-day, 10-day, and 22-day ahead forecasts in the second dataset. Similar to the observations in Tables 4, 5, and 6, the results for long-horizon forecasting in the second dataset are mixed, with no dominant models emerging in forecasting. These results align with the findings from variable selection, highlighting that sentiment variables exhibit a short-term impact, specifically within a 1-day ahead horizon. However, notable improvements persist even in 5 to 22-day ahead forecasting of volatility when incorporating sentiment covariates, particularly evident in the nonlinear XGB model. XGB consistently outperforms random forests and linear models across various forecasting horizons.

Variable Selection	None		Lasso		Adaptive Lasso		Random Forest	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Model 1 LS	**.035**	**.156**	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 LS	.046	.167	**.031**	.150	**.030**	**.146**	**.075**	**.215**
Model 3 LS	.054	.180	.033	**.147**	.033	.149	.078	.228
Model 1 RF	.031	.143	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 RF	.015	**.093**	**.021**	**.11332**	**.011**	**.079**	**.010**	**.078**
Model 3 RF	**.015**	.095	.022	.115	.011	.080	.012	.085
Model 1 XGB	.037	.156	NaN	NaN	NaN	NaN	NaN	NaN
Model 2 XGB	**.010**	**.076**	.024	**.116**	**.008**	**.069**	**.007**	**.065**
Model 3 XGB	.012	.080	**.024**	.118	.008	.070	.008	.070

Table 10: OUT-OF-SAMPLE FORECASTING RESULTS FOR FORECAST HORIZON 22.

5. CONCLUSION

This paper investigates the impact of investor sentiment on stock market volatility using linear and nonlinear models. In particular, we concentrate on forecasting VIX whether sentiment variables, especially sentiment score retrieved from Twitter, is useful in forecasting stock market volatility. Using natural language processing techniques utilizing RoBERTa models, this research computes sentiment scores over 2 million observations of tweets containing certain word like 'Stock Market', 'VIX', and 'Dow Jones'. In contrast to previous literature, this paper considers not only linear-based models for variable selection but also nonlinear tree-based models to assess the robustness of sentiment variables in a nonlinear structure. Furthermore, various cross validation and bootstrapping methodologies are implemented to analyze the time dependent structure of labels and features enhancing the robustness of the sentiment variable's impact.

Our results show that the informativeness of sentimental covariates for future

market volatility is generally strong. In particular, sentiment improves forecasting errors significantly in 1 to 5-day ahead forecasts. We find that these results remain robust in both bagging and boosting algorithms, represented by random forests and XGB. Moreover, the volatility forecasting performances in random forest and XGB models surpass those of linear models with variable selections. In both linear and nonlinear models, the forecasting error predominantly decreases when utilizing sentiment variables in 1/5-day ahead forecasting.

However, as the forecast horizon increases, the effects of sentiment covariates weaken and eventually disappear in 4-week ahead forecasting. The diminishing forecasting power is in line with previous findings on the relationship between investor heterogeneity and volatility, as demonstrated by studies such as Corsi (2009) and Weinbaum (2009), among many others. However, we find that the economic variables retain predictive power up to the 4-week horizon. Additionally, sentiment covariates enhance the predictive accuracy of 10-ahead predictions in random forests and XGB models for market volatility forecasting. In the linear model without variable selections, there is much weaker informativeness of economic/sentiment variables. We cautiously conclude that the complexity between economic/sentiment variables and market volatility can be better modeled by nonlinear models with effective selections, especially in volatility forecasting.

We also consider the fact that sentiment scores extracted from Twitter from January of 2019 seem to have more noise compared to the ones extracted before 2019, show that thorough and relevant filtering processes of twits are necessary to get reasonable conclusion on sentimental variables. Assigning different weights to tweets from users with a large number of followers or giving different weights to tweets with more 'like' buttons can be an interesting extension of our research. Incorporating more machine learning techniques and applications to Korean financial data also are left to future studies.

6. APPENDIX: DATA DESCRIPTION

6.1. SENTIMENT VARIABLES

Consumer sentiment is a statistical measurement of the overall health of the economy, determined by consumer opinion, and widely considered a useful economic indicator. However, the sentiment variables in this paper specifically pertain to the current financial market and consensus, differing somewhat from consumer sentiment.

Data	Description
Score	Daily sentiment data retrieved from Twitter
Google_1	Google search volume data for 'Stock Market'
Google_2	Google search volume data for 'Financial Market'
Google_3	Google search volume data for 'VIX'
Google_4	Google search volume data for 'S&P'
Google_5	Google search volume data for 'MSCI'
Google_6	Google search volume data for 'Dow Jones Industrial Average'

Table 11: SENTIMENT VARIABLES.

6.2. ECONOMIC AND FINANCIAL VARIABLES

Data	Description
GSPC	S&P500 Returns from Yahoo
CL=F	Crude Oil Price from Yahoo
DJI	Dow Jones Industrial Average Returns from Yahoo
MSCI	Morgan Stanley Capital International Returns from Yahoo
Oil	Crude Oil Prices: West Texas Intermediate (WTI)
DXY	Nominal Broad U.S. Dollar Index
Credit	ICE BofA US High Yield Index Option-Adjusted Spread
Term	10-Year Treasury Constant Maturity minus 3-Month Treasury Constant Maturity
FF	Federal Funds Effective Rate (Linear Interpolated)
Mkt-RF	Excess Return on the Market (Fama-French)
SMB	Small Minus Big (Fama-French)
HML	High Minus Low (Fama-French)
ST_Rev	Daily Short-Term Reversal Factor (Fama-French)
Tbill	1-Year Treasury Bill Secondary Market Rate
EUR	U.S. Dollars to Euro Spot Exchange Rate
CNY	Chinese Yuan Renminbi to U.S. Dollar Spot Exchange Rate
GBP	U.S. Dollars to U.K. Pound Sterling Spot Exchange Rate
JPY	Japanese Yen to U.S. Dollar Spot Exchange Rate

Data	Description
GSPC_TR	S&P500 daily volume divided by total market capitalization
DJI_TR	DJI daily volume divided by total market capitalization
MSCI_TR	MSCI daily volume divided by total market capitalization
GSPC_TR_Change	Log-change in the S&P500 turnover-ratio
DJI_TR_Change	Log-change in the DJI turnover-ratio
MSCI_TR_Change	Log-change in the MSCI turnover-ratio
Inf_M	Monthly log change in Consumer Price Index (Interpolated)
SentMichigan	Consumer Sentiment Index from University of Michigan (Interpolated)
M1	Monthly log-change in SA M1 money supply (Interpolated)
House	Monthly log-change in new private housing started (Interpolated)
Indus	Monthly log-change in SA Industrial Production (Interpolated)
Orders	Monthly log-change in SA New Orders (Interpolated)
CRB	Log-return on CRB Index
Capa_Level	Capacity Utilization Level (Interpolated)
Capa_Change	Capacity Utilization Change (Interpolated)

Table 12: ECONOMIC AND FINANCIAL VARIABLES.

REFERENCES

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

Audrino, F., Sigrist, F., and Ballinari, D. (2020). "The impact of sentiment and attention measures on stock market volatility," *International Journal of Forecasting*, 36, 334–357.

Baker, M., and Wurgler, J. (2006). "Investor sentiment and the cross-section of stock returns," *Journal of Finance*, 61, 1645–1680.

Ballestra, L. V., Guizzardi, A., and Palladini, F. (2019). "Forecasting and trading on the VIX futures market: A neural network approach based on open to close returns and coincident indicators," *International Journal of Forecasting*, 35, 1250–1262.

Breiman, L. (2001). "Random forests," *Machine Learning*, 45, 5–32.

Chen, T., and Guestrin, C. (2016). "XGBoost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Christiansen, C., Schmeling, M.,and Schrimpf, A. (2012). "A comprehensive look at financial volatility prediction by economic variables," *Journal of Applied Econometrics*, 27, 956–977.

Cookson, J. A., and Niessner, M. (2019). "Why don't we agree? Evidence from a social network of investors," *Journal of Finance*, 75, 173–228.

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

Fernandes, M., Medeiros, M. C., and Scharth, M. (2014). "Modeling and predicting the CBOE market volatility index," *Journal of Banking & Finance*, 40, 1–10.

Hosker, J., Djurdjevic, S., Nguyen, H., and Slater, R. (2019). "Improving VIX futures forecasts using machine learning methods," *SMU Data Science Review*, 1, 6.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding," *ArXiv: Computation and Language*.

Kim, D., and Baek, C. (2019). "Factor-augmented HAR model improves realized volatility forecasting," *Applied Economics Letters*, 27, 1002–1009.

Kim, B. Y., and Han, H. (2022). "Multi-step-ahead forecasting of the CBOE volatility index in a data-rich environment: Application of random forest with Boruta Algorithm," *Korean Economic Review*, 38, 541-569.

Kim, J., and Baek, C. (2018). "Neural network heterogeneous autoregressive models for realized volatility," *Communications for Statistical Applications and Methods*, 25, 659–671.

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., and Camacho-Collados, J. (2022). "TimeLMs: diachronic language models from twitter," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Politis, D. N., and Romano, J. P. (1994). "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Seo, S. W., and Kim, J. S. (2015). "The information content of option-implied information for volatility forecasting with investor sentiment," *Journal of Banking & Finance*, 50, 106–120.

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Vrontos, S. D., Galakis, J., and Vrontos, I. D. (2021). "Implied volatility directional forecasting: a machine learning approach," *Quantitative Finance*, 21, 1687–1706.

Weinbaum, D. (2009). "Investor heterogeneity, asset pricing and volatility dynamics," *Journal of Economic Dynamics and Control*, 33, 1379–1397.

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, 68, 49–67.

Zou, H. (2006). "The Adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.