# Forecasting Stock Market Volatility: Sentimend Based Approach *

Thomas G. Choi †

December, 2022

**Abstract**

This paper examines the effect of investor sentiment on stock market volatility using natural language processing classification method on a large scale of social network data. This paper also applies numerous forecasting techniques not only including conventional linear models, but also different machine learning and deep learning models and compare its results. Among various economic and sentiment features, lasso-based linear and tree-based non linear variable selection method is performed to show sentiment measures perform a critical role in market volatility. Recalling all the variables are considered as time-series, machine learning based forecasting methodologies are adjusted to capture the time-dependent effect. The results show that sentiment variables are identified to be one of the most important variables in relationship with stock market volatility, and improves future prediction of volatility when considered.

**Keywords:** Forecasting Market Volatility, Investor Sentiment, Machine Learning

*I thank to Changsik Kim, Heejoon Han, and Jihyun Kim for useful comments and discussions.
†31415 Dasan Hall, Sungkyunkwan University, Seoul, Republic of Korea

# 1 Introduction

Market volatility in this paper refers to the VIX index, or implied volatility index of the Chicago Board Options Exchange (CBOE). It is well known that VIX index is designed to measure expectation of 30-day volatility implied by S&P 500 prices of puts and calls and is deduced in real-time. Recalling that VIX index represents expected fluctuations of the market, it is crucial that it informs the market participant about how the market is reacting to future heralded risk. Moreover, as the financial market being more sophisticated with different derivatives emerging with all kinds of underlying assets, VIX futures and options were launched in 2004 and 2006 which delivered efficient information of how participants in financial markets believe. Thus, deliberate predictions of the VIX index will provide valuable information to investors which will alleviate information asymmetry among participants of the market.

Conventional economics maintain that price reflects all available information, but numerous empirical findings suggest anomalies in 'efficient' stock market. This paper mainly studies how investor sentiment, collected from social network service and classified by natural language processing techniques, effect stock market volatility. Over 2.5 million observations of investor sentiment data are collected from Twitter for the most recent 10 years, and were classified by BERT algorithm proposed by Devlin, Chang, Lee, and Toutanova (2018). This paper modifies the heterogeneous auto-regressive (HAR) model from Corsi (2009), not only with using different covariates, but also adapting HAR into different machine learning and deep learning algorithms. Along side with the sentiment data, numerous economic covariates (macroeconomic, fundamental, and technical) are also considered regarding its impact on market volatility.

To clarify whether the sentiment index extracted from social media is actually related and have a significant effect on stock market volatility, lasso-based linear and tree-based non-linear variable selection methods are implemented. The results demonstrate in non-linear case, random forest and XGBoost selects the sentiment feature. In addition, this paper utilize not only the conventional linear regression based methodologies for forecasting future volatility, various machine learning methodologies are also considered. Lasso, ridge, random forest, and XGBoost methods are implemented to determine sentiment variable's impact on future volatility forecasting in both linear and

non-linear perspective.

Results show that economic and sentiment variables plays a crucial role in future volatility forecasting, which coincides with the results from Audrino, Sigrist, and Ballinari (2020), where the authors implemented adaptive lasso method for variable selection and prediction. However, not only in linear-based methodologies, this paper suggests that the relationship still holds when non-linear relationship is premised, demonstrating that sentiment covariate being selected at Random Forest and XGBoost models. Moreover, forecasting stock market volatility with sentiment variables induced lower MSE in both linear models and non-linear models reinforcing the argument that sentiment measures are significant predictors of future market volatility.

# 2 Literature Review

Majorities of studies examined methodologies to accurately forecast market volatility in both specific security based realized volatility and implied volatility based the VIX index. Christiansen, Schmeling, and Schrimpf, (2012) utilized economic variables in order to forecast realized volatility verifying the predictive power of economic variables. Fernandes et al. (2014) employs the HAR model to forecast the VIX index with numerous economic covariates, and also make an attempt to capture the non-linearity relationship by neural network approach. Ballestra et al. (2019) forecasts VIX futures using feed-forward neural network resulting higher accuracy of predictions.

In turn, numerous studies on using investor's sentiment to analyze the behavior of stock market has been growing with bigger data sets 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 (2020) gathers messages from Stocktwits, a social network platform with investors sharing their opinion on different financial securities. The authors classified over 1 million messages and derived the disagreement measure which Antweiler and Frank (2004) proposed, verifying 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 Wulger (2006).

In addition, there has 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. (2018) verified that RNN and 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 play a crucial role in the reduction of prediction errors. Concentrating more on neural network based forecasting, Kim and Baek (2019) demonstrated neural network based HAR model generally performs better than the traditional HAR, but adding too many terms to estimate might decrease its supremacy. Kim and Baek (2020) improves the results by Factor-Augmented HAR model using LSTM networks and showed augmentation of factors improves realized volatility forecasting considering not only S&P, but also numerous stock indices in Asia such as Nikkei, Hangseng, and KOSPI.

This paper mainly follows the scheme of Audrino, Sigrist, and Ballinari (2020), with three 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 non-linear effect of sentiment and economic variables. Moreover, this paper, considering the fact that covariates used are all time-series, implements methodologies to conserve the time-dependent structure of the data when employing cross-validation and bootstrap techniques in machine learning, typically in tree-based models. Remaining parts of this paper is organized as follows. Section 3 provides data and research methodologies, briefly introducing machine learning based techniques. Section 4 summarizes the results of variable selection and forecasting accuracy, and Section 5 concludes.

# 3 Data

Along with CBOE VIX index, this paper considers numerous economic and sentiment variables in order to verify whether investor sentiment plays a crucial role in forecasting, and improve the accuracy of forecast itself. The time span of the data used in this research is from January 1 2010 to December 31 2018, and taking holidays and non-trading days into account, total 2,197 observations are used. Considering the fact that some raw data has different frequencies, interpolation and aggregation methodologies are applied. For forecasting, data until day $t$ in Eastern time are used to predict the VIX index on day $t + 1$.

There are numerous economic variables used in this research to predict the VIX index, most of which are also considered in Audrino et al. (2020) and Kim and Han (2020). It is widely known that economic features related to macro-economy and financial market significantly improves volatility forecasts, not only in realized volatility manners but also in VIX forecasting. According to Audrino et al. (2020) this paper also categorizes macroeconomic variables into five 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, and it's compounded return for 5, 10, 22 days.

- **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 Variable**: 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 Variable**: Inflation Rate (Interpolated), 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.

Along with different economic covariates, this paper also investigates the effect of investor sentiment on the VIX index. This paper considers only Twitter for gathering tweets since Stocktwits limited their API access in 2021. Unlike Audrino et al. (2020) where they used random historical sample from the internet archive, total 2,030,051 numbers of raw tweets are gathered from Twitter itself, using web-scrapping. Comparing with Audrino et al. (2020) where tweets that mentions specific stock are considered for realized volatility of each firms, regarding this research focuses on the overall financial market for VIX forecasting, broad keywords relating with the whole market is selected. Tweets containing words 'S&P500', 'SP500', 'MSCI', 'DowJones', 'DJI', and '$SPY' are scrapped. To filter advertisements and non-related friend chats, tweets containing hyperlinks, tweets which are retweeted, and tweets that repeatedly tweets the same content are all eliminated.

For each tweets, sentiment scores are computed by roBERTa-based model (Loureiro et al. (2022)) from Hugging Face, which is trained for sentiment analysis using Twitter, trained on 124 million tweets. It classifies each tweets and computes 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 it's sentiment scores by assigning scores to each labels, specifically, it assigned 0.727 to negative, 0.256 to neutral, and 0.017 to positive. For computing overall sentiment score for each tweets, 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, deducing one sentiment score for further analysis. The roBERTa model utilizes transformers for it's computation, which is a deep learning model based on neural network. However, unlike traditional RNN and LSTM based machine translation models which process sentence word by word, Transformers are well known to have better performances, with whole sentence being processed and not suffering from long dependency problems.

Sentiment scores, computed for each tweets, are then averaged by each date deducing daily sentiment scores. One noticeable feature from Figure 1 is, variance of daily sentiment score is high during the early periods and gradually decreases as time passes. This is due to the fact that there are significantly less tweets on stock and financial market in the early periods: 24,342 tweets in 2010, 66,098 tweets in 2011, whereas 135,679 tweets in 2018. On the other hand, the high fluctuation in 2018 resembles the stock market crash in Feb 2018, when Dow dropped more than

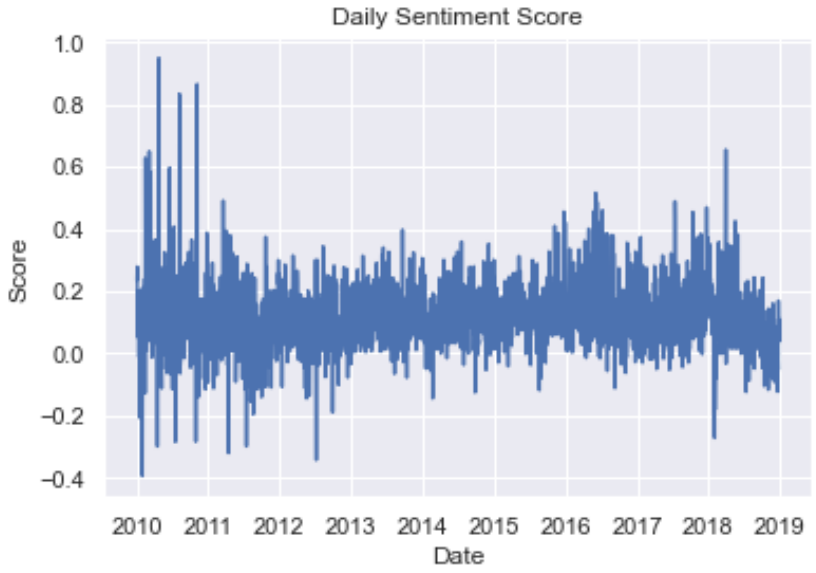


Figure 1: Daily Sentiment Score from 2010 - 2018

12% in two weeks. One can also notice that the mean of sentiment score is over 0, specifically 0.1014. resembling people have usually positive belief on the financial market, coinciding with the results Cookson and Niessner (2020) presented from the messages of Stocktwits.

Along with the sentiment score computed from Twitter, this paper also investigates the effects of different sentiment features, specifically Google search volume data and disagreement score from Cookson and Niessner (2020). Google search volume data retrieved from Google Trends, which only provides relative numbers of daily search queries within 269 days. Thus, this research follows the method of Audrino et. al. (2020) for normalizing the number of Google searches from the beginning of 2010 to end of 2018, with highest day to 100 and lowest day to 0. The Google search volume for words: 'Stock Market', 'Financial Market', 'VIX', 'S&P', 'MSCI', 'Dow Jones Industrial Average' are considered. Although search volumes for 'Financial Market' and 'MSCI' are quite noisy, other covariates tend to capture the movements of financial market itself, with all extremely high search volumes in February 2018.

Last sentiment index this paper takes into account is the disagreement score from Cookson and Niessner (2020). Due to the inaccessibility to Stocktwits data, disagreement score computed by message users of Stocktwits post with bullish or bearish tags are alternatively used. With maximum entropy based classification method, Cookson and Niessner trained the model with messages that

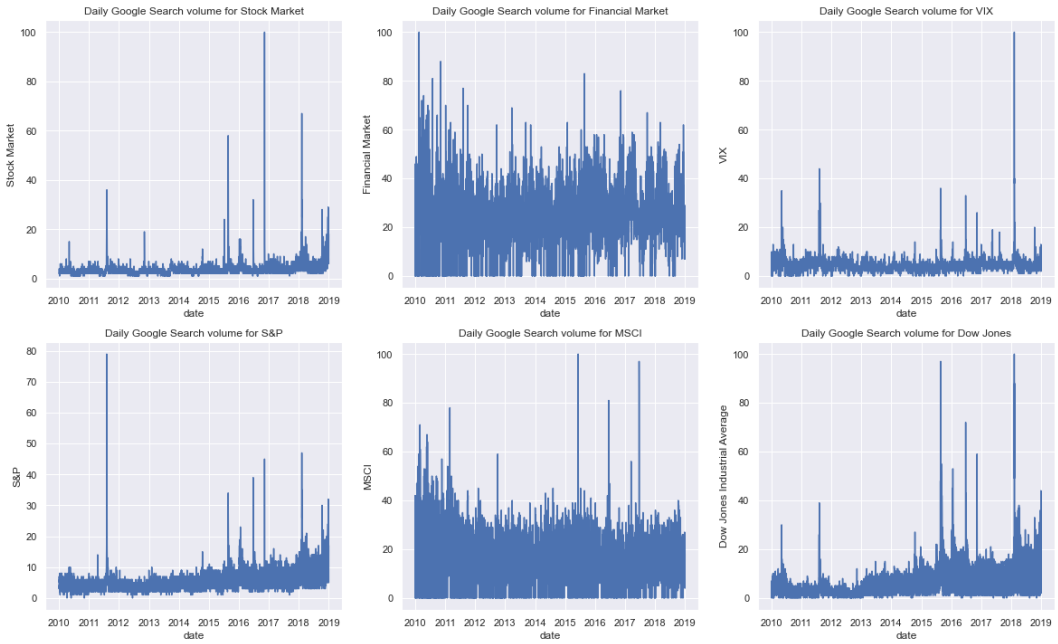


Figure 2: Daily Google Search Volumes from 2010 - 2018

has tags whether indicating it is bullish and bearish, applying it to the rest of unclassified messages. They mention that cross-validation resulted accuracy of 83%. See Cookson and Niessner (2020) for more details. Although the paper only utilizes message posted during Jan 2013 to Sep 2014, their website provides disagreement scores up to Dec 2018.

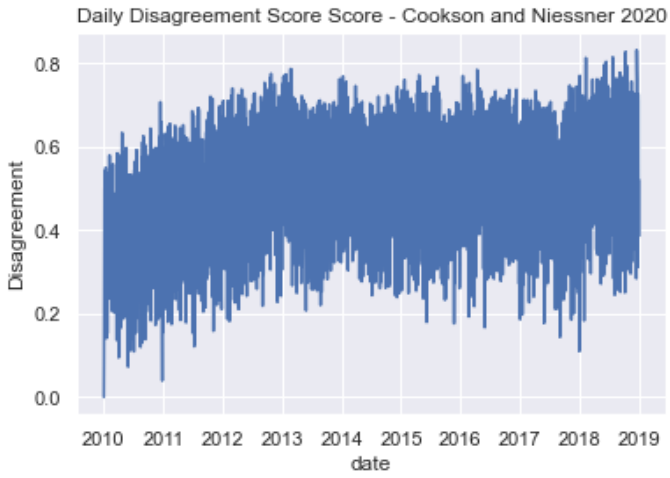


Figure 3: Daily Google Search Volumes from 2010 - 2018

# 4 Results

This section investigates whether economic and sentiment covariates play a significant role in volatility forecasting using linear and non-linear models. With linear models, lasso based models are proposed following the results of Audrino et al. (2020), with variable selection and it's forecasting power comparing benchmark HAR model, economic HAR model, and sentiment & economic HAR model. Taking things further, non-linear models are also proposed on HAR, economic HAR, and sentiment & economic HAR, especially focusing on tree-based models. Ensemble models using bagging (Random Forests) and boosting (XGBoost) are considered, with computation of variable importance and forecasting power.

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 literature already applied HAR models in forecasting VIX; Fernandes et al. (2014), Ballestra et al. (2019), Kim and Han. (2020) are typical examples. The benchmark HAR model in this paper is,

$$\log VIX(D)_{t+1} = \beta_0 + \beta_1 \log VIX(D)_t + \beta_2 \log VIX(W)_t + \beta_3 \log VIX(M)_t + \epsilon_{t+1}$$

where $D$ represent daily, $W$ and $M$ represents weekly and monthly averages of daily log of VIX. The economic HAR model extends the basic HAR model, including equity market, bond market, exchange market, liquidity, and macroeconomic variables, and is represented as,

$$\log VIX(D)_{t+1} = \beta_0 + (\log VIX_t)'\beta_{vix} + M_t'\gamma_{eco} + \epsilon_{t+1}$$

The economic & sentiment HAR model extends the economic HAR model with sentiment variables, specifically Twitter sentiment scores, Disagreement scores, and google search volumes for 6 keywords.

$$\log VIX(D)_{t+1} = \beta_0 + (\log VIX_t)'\beta_{vix} + M_t'\gamma_{eco} + S_t'\theta_{sent} + \epsilon_{t+1}$$

**0. Train and Test Set**

For train and test set, among total of 2,197 observations, first 1,923 observations are used for training and 275 observations are used for testing, with a 90 to 10 ratio of train-test-split. Recalling that for lasso-based linear models, normalization or standardization is essential to match the scale of the penalty term. In addition, normalization is also needed for neural network based models, and thus Min-Max normalization in implemented to all variables. For forecasting and computation of MSE, inverse scaling is implemented so that results can be easily interpreted and compare.

For the train set, two different types of cross validation methods are implemented regarding that all variables are time-series. This paper applies the original k-fold cross validation with 5 folds, which randomly chooses 1 fold as validation set and the other 4 as the train set to adjust hyper-parameters. However, since the series inhere a time dependent structure, there are possibilities that some splits might disintegrate the dependent structure inducing preposterous values for hyper-parameters. Thus, this research also implements time-series split from the Scikit-learn library and consecutively verifies the train and validation sets. For example, if 1st, 2nd, and 3rd block is used for training, then 4th block automatically is used for validation without harming the dependent structure itself.

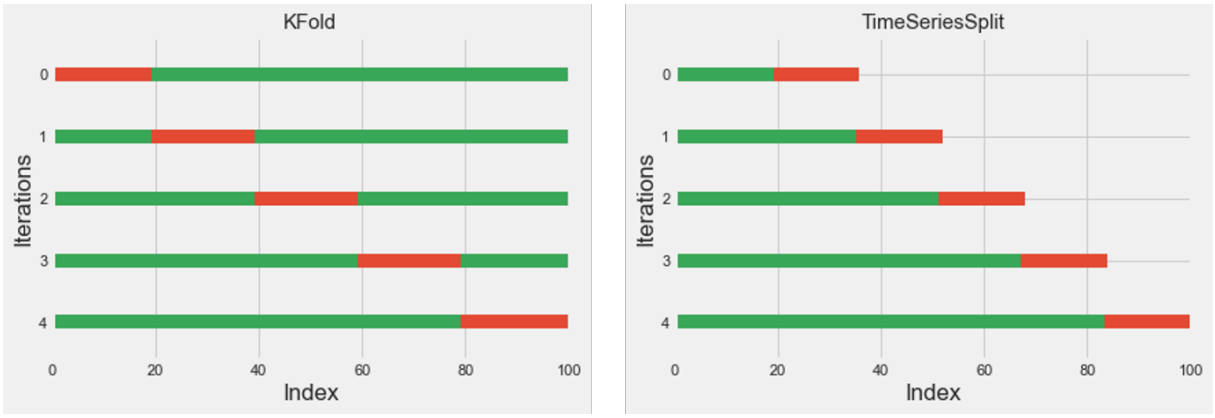


Figure 4: K-Fold and Time-Series Split Cross-Validation

## 1. Linear Models

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, a L1 regularization, only selecting features that have significant impact on the dependent variable by applying a constant penalty term in the original cost function of OLS. $\lambda$, the hyper-parameter, is determined by cross-validation, both the k-fold and time-series split with values ranging from 0.001 to 2 with 0.1 steps, total of 20 different $\lambda$s. Using grid search, both in K-fold and time-series cross-validation, $\lambda$ was selected to be 0.001, and following Audrino et al. (2020) three HAR features (daily, weekly, and monthly) are excluded for variable selection using lasso.

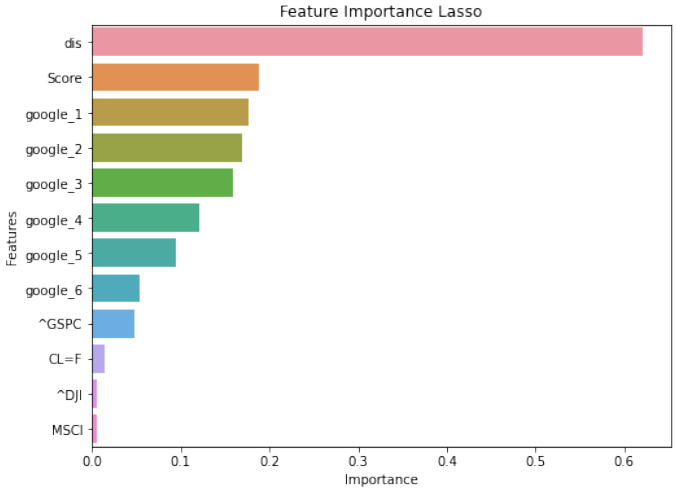


Figure 5: Variable Selection using Lasso

Considering that disagreement and score variable are selected, it is clear that sentiment variables indeed effects stock market volatility. Moreover, one can also see google search volume related with stock market also has a significant impact on volatility. The forecasting results of linear regression, ridge regression, and lasso regression are presented below, with $R^2$, $MSE$, and $MAE$ values. The overall forecast considering economic & sentiment variables tend to have higher forecasting power compared to the benchmark HAR model. For linear regression, $R^2$, MSE, and MAE all implies the **All** Model, indicating economic & sentiment HAR model has the most predictive power. However, ridge and lasso suggests alternative results demonstrating the high predictive power of benchmark HAR in 1-day ahead forecasting, which complies with the results of Kim and Han (2020).

| Model | Linear | Ridge | Lasso |
|---|---|---|---|
| **All $R^2$** | 0.9038 | 0.8770 | 0.8985 |
| **ECO $R^2$** | 0.8990 | 0.8768 | 0.8985 |
| **HAR $R^2$** | 0.9001 | 0.8957 | 0.8991 |
| **All MSE** | 0.0086 | 0.0109 | 0.0090 |
| **ECO MSE** | 0.0090 | 0.0110 | 0.0090 |
| **HAR MSE** | 0.0089 | 0.0093 | 0.0090 |
| **All MAE** | 0.0613 | 0.0703 | 0.0628 |
| **ECO MAE** | 0.0631 | 0.0679 | 0.0628 |
| **HAR MAE** | 0.0622 | 0.0637 | 0.0631 |

Table 1: Forecasting Results of Linear Models

**2. Random Forests**

This paper also investigates the impact of sentiment and economic variables in volatility using non-linear tree-based models. Random Forests, introduced by Brieman (2001), is well known and popular machine learning technique that ensembles numerous decision trees by bagging. Bagging is an algorithm that uses bootstrapping and aggregating individual decision trees to produce stable, non-overfitting results. From the original training set, bagging selects $m$ variables at random from $p$ variables with $m < p$. It randomly chooses $X_i$s, or features and samples uniformly with replacement to create new, or pseudo training sets from the original one. Since the sampling is done with replacement, some observations are repeated, and the sequence of original observations are mixed. For specifics on random forests, please refer to the appendix of this paper.

Numerous economic papers implementing random forests, however, overlook the fact that bootstrapping, or in this case, IID bootstrapping and creating pseudo training sets disintegrates the time-dependent structure that the original sample exhibit. In other words, bootstraps creating by uniform and random sampling doesn't fully represent the original training set's dependency. To solve this problem, this paper implements stationary bootstrap method introduced by Politis and Romano (1994) to substitute the original IID bootstrap method for generating bootstraps. For specifics on stationary bootstraps, please refer to the appendix of this paper.

Recalling the k-fold and time-series split cross-validation methods presented earlier in this paper, random forests also implements two different CV methods for hyper-parameter tuning. For number of estimators (number of trees) are set from 100 to 300 with step set to 50, and max depth (number of splits for each tree) is set from 4 to 14 with step set to 2. With total 30 different parameters for grid search, hyper-parameters of IID bootstrap based random forest and Stationary bootstrap based random forest are separately deduced for forecasting and computing variable importance.

K-fold and time-series split cross-validation results comes out to be very similar implying that k-fold or time-series CV doesn't have a significant effect on selection of best models. For economic and sentiment HAR model, all 4 models, IID Bootstrap with k-fold, IID Bootstrap with time-series split, SB Bootstrap with k-fold, and SB Bootstrap with time-series split all verifies number of estimators to be 150 and max depth to be 4. Economic HAR model had similar results, while benchmark HAR model had 100 number of estimators and max depth set to 6 as the result of cross-validation. This result strengthens the robustness of this paper's results.

The results of forecasting is presented in Table 2. As one can see, max $R^2$ values was derived economic and sentiment HAR model (**All**) from IID Bootstrap with hyper-parameters adjusted from time-series split for cross-validation. In addition, lowest MSE was 0.0086, also derived from economic and sentiment HAR model, regardless of hyper-parameters and bootstrap methodologies. Moreover, the lowest MAE was 0.0629, deduced from economic and sentiment HAR model, in IID Bootstrap with hyper-parameters adjusted from time-series split. Although the bootstrap and cross-validation methodologies didn't have a significant effect on forecasting power, it is distinct that using sentiment and economic variables outperform economic and benchmark HAR models.

In variable importance computed from out-of-bag samples of bagging algorithms, Figure 6, results for economic and sentiment HAR model with IID bootstrap using time-series cross-validation, demonstrates sentiment scores computed from Twitter has the greatest importance after $VIX_t$. $VIX_t$'s importance is omitted due to it's highly significant importance value. In addition, Google search volume for "Dow Jones Industrial Average" was also selected (google_6). The results from feature importance and forecasting verifies that sentiment scores plays a crucial role in improving volatility forecasting, regardless of different hyper-parameters or bootstrapping methods.

| Model | IID RF(K-fold) | IID RF(Time) | SB RF(K-fold) | SB RF(Time) |
|---|---|---|---|---|
| **All $R^2$** | 0.9028 | **0.9034** | 0.9029 | 0.9029 |
| **ECO $R^2$** | 0.8987 | 0.8993 | 0.8989 | 0.8989 |
| **HAR $R^2$** | 0.8912 | 0.8924 | 0.8883 | 0.8980 |
| **All MSE** | **0.0086** | **0.0086** | **0.0086** | **0.0086** |
| **ECO MSE** | 0.0090 | 0.0089 | 0.0090 | 0.0090 |
| **HAR MSE** | 0.0097 | 0.0096 | 0.0099 | 0.0091 |
| **All MAE** | 0.0631 | **0.0629** | 0.0635 | 0.0635 |
| **ECO MAE** | 0.0647 | 0.0646 | 0.0650 | 0.0650 |
| **HAR MAE** | 0.0659 | 0.0655 | 0.0660 | 0.0649 |

Table 2: Forecasting Results of Random Forests

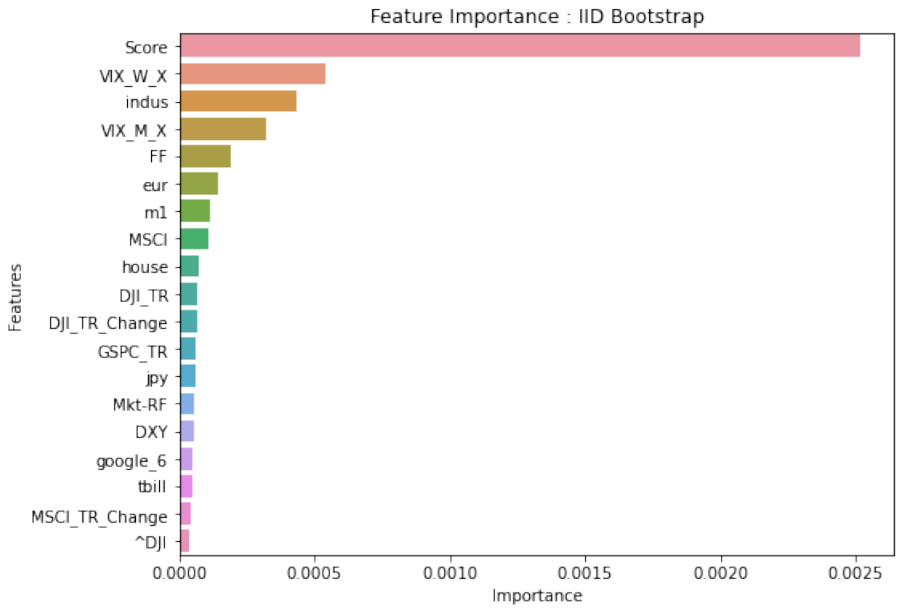


Figure 6: Variable Importance of Random Forests

## 3. XGBoost

Along with bagging algorithms, this paper also investigates the effects of sentiment variables using boosting algorithms, by implementing XGBoost methodology introduced by Chen and Guestrin (2016). Unlike bagging, boosting, especially gradient boosting starts from a weak model and focuses on the cases where the former model had problems solving it. Usually it starts from a stump tree, a decision tree which only uses one attribute for splitting. Considering from a regression perspective, the second train set is determined by the true values of $y$ minus the fitted values of $y$, or $\hat{y}$, which is the residuals. This process is repeated until it reaches a point where training errors does not decrease anymore. Recalling that the gradient of loss function is the fitted value minus true $y$, the algorithm is called Gradient Descent since it moves to the opposite direction of the gradient itself. For more specifics, please see the appendix of this paper.

| **Model** | **XGB (K-fold)** | **XGB (Time)** |
|---|---|---|
| **All** $R^2$ | 0.8673 | **0.8777** |
| **ECO** $R^2$ | 0.8688 | 0.8217 |
| **HAR** $R^2$ | 0.8761 | 0.8761 |
| **All MSE** | 0.0118 | **0.0109** |
| **ECO MSE** | 0.0117 | 0.0159 |
| **HAR MSE** | 0.0110 | 0.0110 |
| **All MAE** | 0.0808 | 0.0740 |
| **ECO MAE** | 0.0736 | 0.0945 |
| **HAR MAE** | 0.0688 | 0.0688 |

Table 3: Forecasting Results of XGBoost

The forecasting results of XGBoost are presented below on Table 3. Similar to random forests, cross-validation methodologies didn't have a significant effect on hyper-parameter tuning, both k-fold and time-series split with 100 estimators and max depth set to 4 induced the lowest MSE in training set. Although the lowest MAE values for XGBoost was derived from the benchmark HAR model, one can easily see that highest $R^2$ and lowest MSE was derived from economic and sentiment

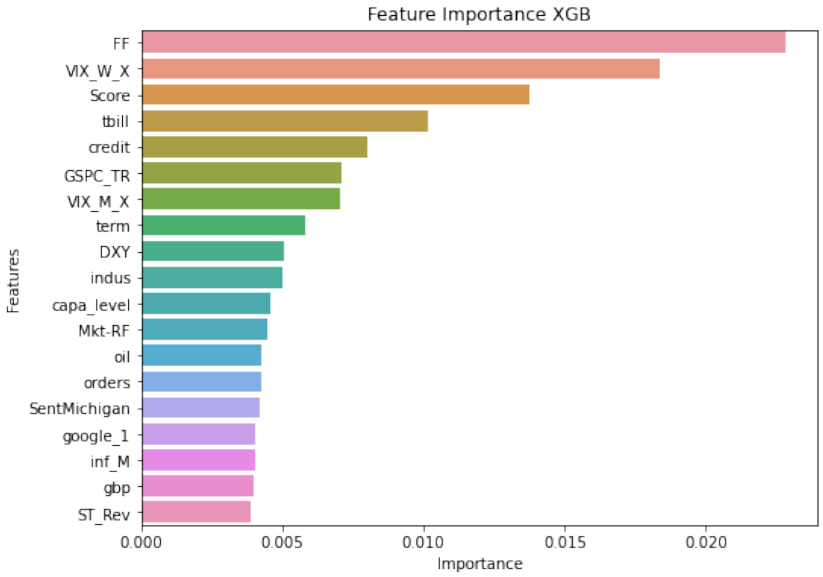


Figure 7: Variable Importance of XGBoost

models. However, compared to the results of random forests, the difference between benchmark HAR model wasn't significantly less, implying that in boosting methodologies, the benchmark HAR could perform better in other time-specific data.

Figure 7 demonstrates the feature importance results from XGBoost. $VIX_t$'s importance is also omitted due to it's significance, and although sentiment score variable isn't as important as in random forests, it still is rated top 3 among economic and sentiment variables. With the results of variable importance and MSE from forecasting, it is clear that sentiment score plays an important role in volatility forecasting.

# 5 Conclusion

This paper investigates the impact of investor sentiment to stock market volatility using linear and non-linear models. With natural language processing techniques utilizing roBERTa models, this research computes sentiment scores over 2 million observations of tweets containing certain word such as 'Stock Market', 'VIX', and 'Dow Jones'. Unlike previous literature, this paper not only considers linear-based models for variable selection, but also non-linear, tree-based models to check the robustness of sentiment variables in non-linear structure. Furthermore, different cross-validation

and bootstrapping methodologies are implemented to analyze the time-dependent structure of label and features, providing robustness the sentiment variable's impact.

Results show that not only in linear models, but also in non-linear models, sentiment improves forecasting errors regarding two types of cross-validation, k-fold and time-series split. In addition, the results stay robust in both bagging and boosting algorithms, represented by random forests and XGBoost, with random forest outperforming XGBoost in reducing forecasting errors. Although the forecasting results of linear and non-linear models are almost same, variable importance from tree-based models selected sentiment score as one of the highest impact factors on stock market volatility, while linear based lasso failed to select and resulted a poor performance.

This paper only concentrates on 1-day ahead forecasting, but further research could show the forecasting results of 5-day or 22-day ahead forecasting to verify the effects of sentiment variable. In addition, sentiment index computed not only from Twitter, but from other sources, such as Stocktwits, could improve the forecasting results, regarding that Twitter is for general users including the ones who does not invest, but Stocktwits have high quality data with different investors and experts. Heterogeneity of investors regarding their perspective in methodologies could be a interesting topic, as Cookson and Niessner (2020) demonstrated.

# References


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. https://doi.org/10.1111/j.1540-6261.2004.00662.x

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

BAKER, M., & WURGLER, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Ballestra, L. V., Guizzardi, A., & 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*(4), 1250–1262. https://doi.org/10.1016/j.ijforecast.2019.03.022

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Christiansen, C., Schmeling, M., & Schrimpf, A. (2012). A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, *27*(6), 956–977. https://doi.org/10.1002/jae.2298

COOKSON, J. A., & NIESSNER, M. (2019). Why Don’t We Agree? Evidence from a Social Network of Investors. *The Journal of Finance, 75*(1), 173–228. https://doi.org/10.1111/jofi.12852

Corsi, F. (2008). A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, *7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

Fernandes, M., Medeiros, M. C., & Scharth, M. (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking &Amp; Finance*, *40*, 1–10. https://doi.org/10.1016/j.jbankfin.2013.11.004

Hosker, J. (n.d.). *Improving VIX Futures Forecasts using Machine Learning Methods*. SMU Scholar. https://scholar.smu.edu/datasciencereview/vol1/iss4/6/

Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv: Computation and Language*.

*Kenneth R. French - Data Library*. (n.d.). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Kim, D., & Baek, C. (2019). Factor-augmented HAR model improves realized volatility forecasting. *Applied Economics Letters*, *27*(12), 1002–1009. https://doi.org/10.1080/13504851.2019.1657554

Kim, & Han. (2022). Multi-Step-Ahead Forecasting of the CBOE Volatility Index in a Data-Rich Environment: Application of Random Forest with Boruta Algorithm. *The Korean Economic Review*, *38*(3). https://doi.org/10.22841/kerdoi.2022.38.3.007

Kim, J., & Baek, C. (2018). Neural network heterogeneous autoregressive models for realized volatility. *Communications for Statistical Applications and Methods*, *25*(6), 659–671. https://doi.org/10.29220/csam.2018.25.6.659

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., & Camacho-collados, J. (2022). TimeLMs: Diachronic Language Models from Twitter. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.18653/v1/2022.acl-demo.25

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, *99*(1), 39–61. https://doi.org/10.1016/s0304-4076(00)00030-0

Seo, S. W., & Kim, J. S. (2015). The information content of option-implied information for volatility forecasting with investor sentiment. *Journal of Banking &Amp; Finance*, *50*, 106–120. https://doi.org/10.1016/j.jbankfin.2014.09.010

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: a machine learning approach. *Quantitative Finance*, *21*(10), 1687–1706. https://doi.org/10.1080/14697688.2021.1905869