SOAS University of London

Centre for Global Finance

Working Paper Series

No.1 / 2021

Forecasting the CBOE VIX with a hybrid LSTM-ARIMA model and sentiment analysis

By Yossi Shvimer, Victor Murinde and Avi Herbon

The Centre for Global Finance (CGF) Working Paper Series features recent studies by resident members of CGF as well as visiting researchers, altogether demonstrating the depth and breadth of research being undertaken at CGF. The papers are published to facilitate preliminary dissemination of ongoing research, enhance quality of work and contribute to the advancement of knowledge. We acknowledge, without implication, financial support from the DEGRP Research Grant (ES/N013344/2) on "Delivering Inclusive Financial Development and Growth", funded by the ESRC and the former UK Department for International Development, which merged with the Foreign & Commonwealth Office on 2 September 2020 to become the Foreign, Commonwealth & Development Office (FCDO), the ESRC-NSFC (ES/P005241/1) Research Grant on "Developing financial systems to support sustainable growth in China – The role of innovation, diversity and financial regulation", and the AXA Research Fund.

List of previous Working Papers of CGF:

No.1/2018 *Capital, risk and profitability of WAEMU banks: Does cross-border banking matter?* By Désiré Kanga, Victor Murinde, and Issouf Soumaré

No.2/2018 *Capital flows and productivity in Africa: The angel is in the details.* By François A. B. Bationo, Stephany Griffith-Jones, Victor Murinde, Issouf Soumaré and Judith Tyson

No.3/2018 *The persistence of bank fragility in Africa: GMM dynamic panel data evidence.* By Abbi M. Kedir, Syed Faizan Iftikhar, Victor Murinde and Bernadette Kamgnia

No.4/2018 *Reflections on central banking*. By Victor Murinde and Patrick Njoroge

No.5/2018 *Let beholders behold: Can banks see beyond oil booms and mitigate the Dutch disease?* By Morakinyo O. Adetutu, John E. Ebireri, Victor Murinde and Kayode A. Odusanya

No.6/2018 *National culture, CEO power and risk-taking by global banks.* By Eilnaz Kashefi Pour and Victor Murinde

No.7/2018 *Corporate investment, financing and payout decisions under financial constraints and uncertainty: Evidence from UK firms*. By Qingwei Meng, Victor Murinde and Ping Wang

No.8/2018 *Bank opacity and risk-taking: Evidence from analysts' forecasts* By Samuel Fosu, Collins G. Ntim, William Coffie, and Victor Murinde

No.9/2018 *Does microcredit increase hope, aspirations and well-being? Evidence from Sierra Leone.* By Adriana Garcia, Robert Lensink, and Maarten Voors

No.10/2018 *Lessons from emerging economies for African low income countries on managing capital flows.* By Stephany Griffith-Jones and José Antonio Ocampo

No.11/2018 *Financial inclusion and economic growth: What do we know?* By Joshua Y. Abor, Haruna Issahaku, Mohammed Amidu, and Victor Murinde

No.12/2018 *Climate vulnerability and the cost of debt.* By Gerhard Kling, Yuen C Lo, Victor Murinde, and Ulrich Volz

No.13/2018 *Pan-African banks on the rise: Does cross-border banking increase firms' Access to finance in WAEMU?* By Désiré Kanga,Victor Murinde, Lemma Senbet, and Issouf Soumaré

No.14/2018 *The peer monitoring role of the interbank market and implications for bank regulation: Evidence from Kenya.* By Victor Murinde, Ye Bai, Christopher J. Green, Isaya Maana, Samuel Tiriongo, and Kethi Ngoka-Kisinguh

No.1/2019 *Central bank independence: What are the key issues*? By Désiré Kanga and Victor Murinde

No.2/2019 *Banking services and inclusive development in sub-Saharan Africa.* By Haruna Issahaku, Mohammed Amidu and Aisha Mohammed Sissy

No.3/2019 *A survey of literature on financial literacy and financial behaviour: Is there a gender gap*? By Maryam Sholevar and Laurence Harris

No.4/2019 *Capital adjustment over the cycle: Evidence from microfinance institutions*. By Issouf Soumaré, Hubert Tchakoute Tchuigoua, and Hélyoth T.S. Hessou

No.5/2019 *Provisioning and business cycle: Evidence from microfinance institutions*. By Hélyoth T.S. Hessou, Robert Lensink, Issouf Soumaré, and Hubert Tchakoute Tchuigoua

No.6/2019 *Lending and business cycle: evidence from microfinance institutions*. By Hubert Tchakoute Tchuigoua, Issouf Soumaré, and Hélyoth T.S. Hessou

No.7/2019 *Term structure of CDS spreads & risk-based capital of the protection seller: an extension of the dynamic Nelson-Siegel model with regime switching.* By Standley R. Baron and Issouf Soumaré

No.8/2019 *Confidence, financial inclusion and sustainable economic development.* By Ayse Demir, Reinhard Bachmann, Victor Murinde, Laurence Harris, Christine Oughton and Meng Xie

No.9/2019 *The network structure of the Malawi interbank market: implications for liquidity distribution and contagion around the banking system.* By Esmie Koriheya Kanyumbu

No.10/2019 *Aid and Exchange Rates in sub-Saharan Africa: No More Dutch Disease?* By Oliver Morrissey, Lionel Roger and Lars Spreng

No.11/2019 *Does credit deepening increase financial fragility*? By Peng Yiqing, Niels Hermes, and Robert Lensink

No.12/2019 *Does microcredit increase aspirational hope? Evidence from a group lending scheme in Sierra Leone.* By Adriana Garcia, Robert Lensink, and Maarten Voors

No.13/2019 *Do better formal institutions promote financial inclusion*? By Peng Yiqing, Niels Hermes, and Robert Lensink

No.14/2019 *Do interbank interest rates reflect the financial soundness of borrowing banks*? By Thomas Bwire, Martin Brownbridge, Doreen K. Rubatsimbira and Grace A. Tinyinondi

No.15/2019 *Institutional environment and the microstructure of the interbank market*. By Thomas Bwire, Martin Brownbridge, Doreen K. Rubatsimbira, and Grace A. Tinyinondi

No.16/2019 *Segmentation of the interbank money market in Zambia*. By Jonathan M Chipili, Francis Z Mbao, Alick B Lungu, Shula M Sikaona, Anthony Bwalya, and Cosam S Chanda

No.1/2020 *How has the rise of Pan-African banks impacted bank stability in WAEMU?* By Désiré Kanga, Victor Murinde, and Issouf Soumaré

No.2/2020 *Threshold effects of financial inclusion on income inequality.* By Ayse Demir, Vanesa Pesqué-Cela, and Victor Murinde

No.3/2020 *FinTech, financial inclusion and income inequality: A quantile regression approach.* By Ayse Demir, Vanesa Pesqué-Cela, Yener Altunbas, Victor Murinde

No.4/2020 *Director reputation and earnings management: evidence from the British honours system*. By Tolulola Lawal

No.5/2020 *Financial inclusion and the growth-inequality-poverty triangle: New evidence from Africa.* By Ayse Demir and Victor Murinde

No.6/2020 Fellowship amongst peers: A systematic and selective survey of literature on the analysis of interbank lending networks. By Anosi F. Ikimalo and Victor Murinde

No.7/2020 *Exploring the impact of COVID-19 pandemic on Africa's FinTech space.* By Joshua Yindenaba Abor

No.8/2020 *Financial market integration in sub-saharan Africa: How important is contagion*? By Robert Akunga, Ahmad Hassan Ahmad and Simeon Coleman

No.9/2020 *Finance and well-being in developing countries: Does access to mobile money improve household well-being*? By Fei Jiang, Christopher J. Green, Ahmad Hassan Ahmad and Carlos Sakyi-Nyarko

No.10/2020 *Mobile money, ICT, financial inclusion and inclusive growth: How different is Africa*? By By Fei Jiang, Christopher J. Green, Ahmad Hassan Ahmad and Victor Murinde

No.11/2020 *Financial inclusion and welfare improvement: Empirical evidence from A households survey data*. By Carlos Sakyi-Nyarko, Ahmad Hassan Ahmad and Christopher J. Green

No.12/2020 *The FinTech revolution: What are the opportunities and risks*? By Victor Murinde and Efthymios Rizopoulos

No.13/2020 *The COVID-19 pandemic and its impact on African economies and financial markets: A review*. By Elikplimi Komla Agbloyor and Joshua Yindenaba Abor

No.14/2020 *Online data collection for monitoring the impact of COVID-19*. By Victor Murinde, Athina Petropoulou and Meng Xie

No.15/2020 *Towards resolving the partner selection problem in venture capital syndication: new insights from a neural network based study*. By Qiong Ji, Xiaoming Ding and Victor Murinde

No.16/2020 *Government policy and financial inclusion: Analysing the impact of the Indian national mission for financial inclusion*. By Rachel Hadar and Ronny Manos

All CGF Working Papers can be downloaded from CGF Website.

Centre for Global Finance
SOAS University of London
10 Thornhaugh Street, Russell Square
London
WC1H 0XG

Email: cgf@soas.ac.uk
Website: https://www.soas.ac.uk/centreforglobalfinance/publications/

Forecasting the CBOE VIX with a hybrid LSTM-ARIMA model and sentiment analysis*

Yossi Shvimer, Centre for Global Finance, SOAS University of London, UK.
ys19@soas.ac.uk

Victor Murinde, Centre for Global Finance, SOAS University of London, UK.
V.Murinde@soas.ac.uk

Avi Herbon, Department of Management, Bar-Ilan University, Israel.
avher@bezeqint.net

Abstract

Forecasting stock market behavior is a challenging problem. Forecasting the risk level associated with stock price futures provides an additional complexity level since it represents a derivative (hence, less stable, with more degree of freedom) of the stock market dynamics. This paper introduces a new next day forecasting model for the Chicago Board Options Exchange (CBOE) Volatility Index (VIX). We show the advantages of adding investors' sentiment scores to a new hybrid model encompassing the long-short term methodology (LSTM) model and the auto-regressive integrated moving average (ARIMA) model. The sentiment scores are empirically evaluated via machine learning natural language processing (NLP) tools based on commonly used economic sites. The hybrid model shows robust results on the forecasts of the next day VIX level. Based on out-of-sample for 2019-2020 end of day data (including COVID-19 out-of-sample data), the hybrid LSTM-ARIMA model shows that the addition of sentiment scores yields higher accuracy by 5% than the hybrid LSTM-ARIMA without investors' sentiment scores, consequently generating higher trade profits.

Keywords: Forecasting; VIX index; Hybrid LSTM-ARIMA model; COVID-19 crisis; Sentiment analysis

JEL Classification: C55

* We acknowledge research funding for the SOAS Centre for Global Finance, under: the AXA Research Fund; the ESRC and the former UK Department for International Development, which merged with the Foreign & Commonwealth Office on 2 September 2020 to become the Foreign, Commonwealth & Development Office (FCDO), Research Grant No. ES/N013344/2 on "Delivering Inclusive Financial Development and Growth". Murinde also acknowledges support under the ESRC-NSFC (ES/P005241/1) Research Grant on "Developing financial systems to support sustainable growth in China – The role of innovation, diversity and financial regulation".

1. INTRODUCTION

Stock market forecasting is considered an essential theoretical as well as practical problem in economics. Timely prediction of market dynamics is challenging due to volatile stock market characteristics (Jin et al., 2019). Many efforts to forecast stock market volatility are recorded in existing literature (see Liu et al., 2015; Lalancette and Simonato, 2017). Yet, the results of these efforts have been rather diverse (Taylor, 2019).

The Chicago Board Options Exchange (CBOE) Volatility Index (VIX) measures the expected volatility associated with the S&P 500 index returns over the subsequent 30 days, as implied by the prices of the basket of options contracts (on the S&P 500 index) with maturities between 23 and 37 days. Given the nature of the underlying asset (S&P500 index), VIX futures contracts are cash-settled, with final settlement taking place on Wednesday, which is 30 days before the third Friday of the subsequent expiry month. These futures contracts' primary purpose is to enable hedgers and speculators to trade volatility at a low cost (high liquidity) environment. Forecasting VIX level has received lesser attention than stock indices by existing literature in this regard (Psaradellis and Sermpinis, 2016), although the VIX has become the standard benchmark for measuring stock market volatility for S&P500 (Qiao et al., 2020). Various previous pricing studies about the forecasting VIX index level used the generalized autoregressive conditional heteroskedasticity (GARCH)-type models. GARCH models try to minimize the noise from the time-series data itself. The basic GARCH specification captures time-varying volatility and incorporates known characteristics of real-world return processes, including asymmetric response to up and down shocks as well as jumps that are well suited for VIX estimation. Wang et al. (2017) estimated VIX futures contract prices based on the Heston and Nandi GARCH (HN-GARCH) model. They show that the VIX index level and VIX futures prices for a joint estimation can effectively capture the variations of the market VIX index level and the VIX futures contract prices simultaneously. Guo and Liu (2020) studied the out-of-sample VIX futures pricing based on GARCH and Goldstein-Jagannathan-Runkle GARCH (GJR-GARCH) models. They found that concerning pricing errors of the VIX futures contracts prices and the VIX index's level, the new methods significantly outperform a continuous-time benchmark based on the Heston volatility model (Heston, 1993). Yang and

Wang (2018) found that the IG-GARCH model can reduce the absolute pricing errors in evaluating the VIX index level obtained by the HN-GARCH model by 11–29%.

Financial time series, such as stock prices and VIX index level, are vulnerable to behavioral factors such as risk aversion and exogenous factors such as macroeconomic shocks. Both elements are practically impossible to capture with existing mathematical models and add noise to time series estimations. Linear models (like those that dominate the relevant literature) are only partially successful in capturing the relevant underlying trend (Psaradellis and Sermpinis, 2016). These models have low forecasting accuracy and high volatility (LeBaron, 2000; Qi and Wu, 2006). In particular, the models seem to be of limited assistance to traders in terms of generating profitable trades.

Recent studies reported that news articles could improve the accuracy of predicting stock price movements. For example, Xu and Cohen (2018) examined the effectiveness of deep generative approaches for stock movement prediction from social media data using a neural network architecture for this task. They tested their model on a new comprehensive dataset and showed that their model increased price accuracy rather than without using social media. Yang et al. (2018) introduced a knowledge-based method to extract relevant financial news adaptively. They used an output attention mechanism to allocate different weights to different days to stock price movement. Through empirical studies based upon three individual stocks' historical prices, they showed an accuracy of 68%, higher than 58% accuracy obtained by sentence embeddings input and standard neural network prediction model. These studies showed that news articles could improve accuracy in predicting stock price movement. However, none of the models predicting the VIX index level involved higher volatility than other underlying assets.

We draw insights from machine learning framework by using the long-short term methodology (LSTM) from Hochreiter and Schmidhuber (1997) and the Auto-Regressive Integrated Moving Average (ARIMA) model (Koreisha and Fang, 1999). Both models are known for processing and forecasting values based on time-series data. ARIMA models assume that the present data have a linear function of past data points and past errors. These errors are white in nature and require that the data be made stationary before fitting a linear equation to the data.

Because both LSTM and ARIMA models can forecast time-series, several papers empirically compare these prediction models. Siami-Namini et al. (2018)

empirically compared ARIMA with LSTM prediction models for several stock indices between 1985 and 2018. They found that LSTM outperforms traditional-based algorithms such as the ARIMA model. The root-mean-square error rates (%) obtained by ARIMA was 55.3 while the correspond value from LSTM was 7.814, indicating the superiority of LSTM for the given dataset. Therefore, models tried to hybridize ARIMA and LSTM models to improve the forecast accuracy of either of the models used separately. For example, Zhang (2003) proposed a hybrid methodology that combines both ARIMA and Artificial Neural Network (ANN) models and showed empirically on real datasets of Forex indicator that the integrated model can improve the forecast accuracy of either of the models used separately. Khashei and Bijari (2011) used ARIMA models to identify and magnify the existing linear structure in data, and then a multilayer perceptron to determine a model to capture the underlying data generating process and predict the future price, using preprocessed data. Khashei and Bijari (2011) showed that their model had better performance for one-step-ahead performance than Zhang (2003). Babu and Reddy (2014) used the Hybrid ARIMA-LSTM methodology by filtering the data using the moving average for the trend. Then they estimated the trend with the ARIMA model and with the LSTM model, the noise from the trend separately. Their one-step-ahead forecast had higher accuracy than both Zhang (2003) model and the Khashei and Bijari (2011) model.

Our investigation extends the above work of Babu and Reddy (2014) and follows the growing literature of using machine learning to forecast the VIX index level. We introduce a new forecasting model of VIX level using the LSTM algorithm and ARIMA model and validate it for real-time data. In particular, we propose a new hybrid LSTM-ARIMA model, with deep learning, for forecasting VIX, and add a new feature, the investors' emotional tendency to financial news. Our main contributions are threefold. First, we suggest engaging investors' sentiment using natural language processing (NLP), which relies on machine learning techniques to parsing text sentiment obtained from financial news articles for stock prediction. Investors' sentiment can improve prediction accuracy obtained by models basing their predictions only on historical prices. Second, to better forecast future values, we add to the LSTM model the investors' sentiment. The LSTM approach has the advantages of analyzing relationships among time-series data through its memory function. This property can quantify the long-term relationship between sentiment analysis data and

VIX values. Third, we adopt the ARIMA model and combine it with the LSTM method to capture the VIX index's trend fluctuation.

The remainder of this article is structured as follows. In Section 2, we explain the data, software, and hardware of our proceeding analysis. The methodology, which consists of three stages, is discussed in Section 3. Section 4 presents the key results of applying the suggested model and some robustness tests. The conclusion is offered in Section 5.

2. DATA, SOFTWARE, AND HARDWARE

2.1. VIX Index Data

We use the S&P 500 VIX index data from CBOE for the empirical study. The entire dataset is from September 30th, 2016, until October 30th 2020, covering 1027 trading days, consisting of the COVID-19 period characterized by a sharp increase in the VIX index level.

2.2. Financial websites data

Using Webhose.io platform published from October 30th, 2016 until September 30th, 2020, we have collected from major financial websites mentioning the word "*S&P500*". These websites include CNN, Reuters, Bloomberg, etc. The cause for using the keyword "*S&P500*" rather than the "*VIX*" keyword itself draws from the assumption under which the investor sentiment of the underlying asset (S&P500) affects the VIX index level. The total number of financial news articles in the whole dataset is 73,566.

2.3. Software and hardware

We used Python 3.7 (Python Software Foundation, 2016), with *NumPy* (Van Der Walt et al., 2011) and *pandas* (McKinney, 2010) packages for data preparation. We developed the architecture of deep learning LSTM networks with *Keras* (Chollet, 2015) on top of Google *TensorFlow*, a powerful library for large-scale machine learning on heterogeneous systems (Abadi et al., 2016). Investors' sentiment analysis uses NLP methods and algorithms that are either rule-based, hybrid, or rely on machine learning techniques to learn data from datasets. The investors' sentiment analysis in our study is carried out separately with TextBlob Library (Loria, 2018) as

well as with Valence Aware Dictionary for Sentiment Reasoning ("*Vader*") that construct and empirically validate a list of lexical features (Hutto and Gilbert, 2014). Vader is a human-validated sentiment analysis method developed for micro-blogging and social media, requiring no training data. It consists of a list of lexical features and associated sentiment measures. Based on the language's grammatical and syntactic usage, several rules are formed, which are used to determine the text's sentiment. A vocabulary, whereas each word is assigned to a semantic orientation as a positive or negative value (Urologin, 2018).

3. HYBRID LSTM-ARIMA MODEL

3.1 The general framework

Our methodology consists of three stages shown in Figure 1. First, we build the input vector based on historical VIX index levels and investments' sentiment analysis scores necessary for training and forecasting VIX index level. Second, we provide an in-sample analysis separately for LSTM networks and the ARIMA model based on 70% of the data ("train data"). We present the Hybrid LSTM-ARIMA approach to obtain forecasts, based on the advantages of ARIMA and LSTM models. Using the Hybrid LSTM-ARIMA model, our goal is to estimate the VIX index level in day $t+1$. The rest of this section details the three stages outlined above and are illustrated in Figure 1. Third, we make out-of-sample data forecasting, separately for the LSTM model and ARIMA model based on the remaining 30% of the data ("trade data").

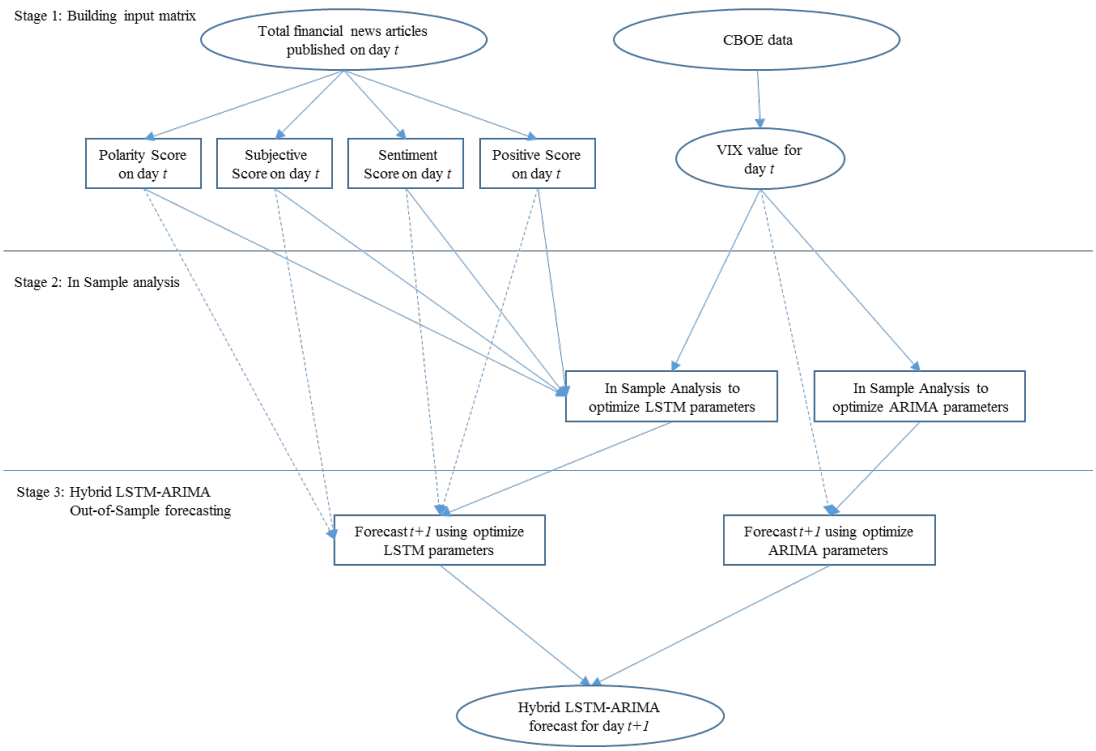

Figure 1. The suggested Hybrid LSTM-ARIMA methodology for forecasting VIX index level in t+1

3.2 Input vector and target generation

LSTM networks and the ARIMA models require time series of input data for training, i.e., the values at successive points in time. Our modeling method receives the VIX index's historical levels and investors' sentiment scoring as key inputs.

3.2.1 Calculating investors' sentiment score

We used the TextBlob library in *Python* to compute the sentiment (polarity) score for each news article out of the $N = 73,566$ news articles along the 1027 trading days. Let n_t denote the total number of articles published in a day t. In TextBlob, the polarity score $tp_{i,t}$ for the given article text of index $i = 1,2,...n_t$=in a day t is in the range of [-1,1]. If the polarity score is positive, it is regarded as positive sentiment, meaning that the news article is positive in the sense of semantic total positive words. If the polarity score is negative, it is regarded as a negative sentiment, meaning that the news article is negative in the sense of semantic total positive words. If the polarity is equal or close to zero, it is considered neutral. We use the TextBlob library

also to compute the subjectivity score $ts_{i,t}$ for given article text i on day t. The subjectivity score $ts_{i,t}$ is within the range [0,1], where 0.0 is a very objective article text, and 1.0 is a very subjective article text. In addition, we use the Vader Sentiment Analyzer library in Python to calculate for given article text i on day t the sentiment score $vs_{i,t}$ of each news article. The Vader sentiment score $vs_{i,t}$ is calculated by summing each word's valence scores in the lexicon and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). We also use Vader positive score $vp_{i,t}$ from the Vader Sentiment Analyzer library for a given article text index i, which scales the *intensity* on a scale between -4 (extremely negative) and 4 (extremely positive).

We compute, for each input individually the daily scores p_t, s_t, v_t, b_t on day t, which is the arithmetic average score of all n_t article text scores as follows:

polarity score $TP_t = \sum_{i=1}^{n_t} tp_{i,t} / n_t$, subjectivity score $TS_t = \sum_{i=1}^{n_t} ts_{i,t} / n_t$, sentiment score $VS_t = \sum_{i=1}^{n_t} vs_{i,t} / n_t$, and Vader positive score $VP_t = \sum_{i=1}^{n_t} vp_{i,t} / n_t$. Table 1 presents the statistics for each sentiment score for the total dataset. We calculated the scores based on the daily average score.

Table 1. Summary of total sentiment scores for the in-sample dataset

Sentiment Score Source	Average Score	Median Score	Max Score	Min Score	Std. Dev.
TextBlob Polarity (TP)	0.077	0.078	0.733	-0.265	0.063
TextBlob Subjectivity (TS)	0.395	0.405	0.900	0.000	0.092
Vader Sentiment (VS)	0.530	0.963	1.000	-1.000	0.732
Vader Positive (VP)	0.088	0.085	0.271	0.000	0.034

Table 1 shows that the sentiment score source (TextBlob polarity, TextBob subjectivity, Vader sentiment, and Vader positive) and the corresponding average scores, median score maximum score, minimum score and the standard deviation of the scores. It is shown that the average and median scores are higher than their mid-range values, meaning that in general, the data relatively positive.

While TextBlob polarity shifts in relatively narrow boundaries, the Vader sentiment score is more diverse, indicating a higher sensitivity for each article's text lexicon. The average subjectivity score suggests that the news contains more objective data than subjective data.

Figures 2a-2d present the correlation between different sentiment scores and VIX level. Each dot in Figures 2a-2d characterizes the TextBlob polarity (TP), TextBob subjectivity (TS), Vader sentiment (VS), and Vader positive (VP) sentiment scores on day *t* (y-axis) and VIX level on day *t* (x-axis) for the in-sample data (718 trading days), respectively.

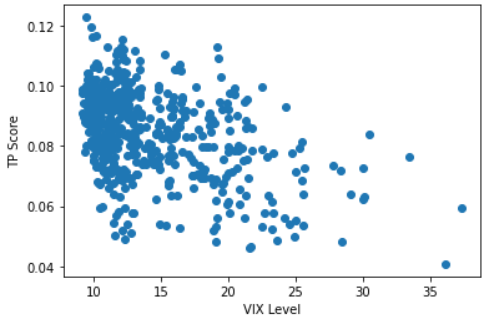

Fig 2a. VIX and TP (correlation= -0.42)

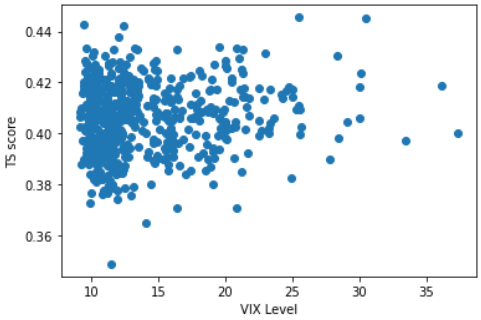

Fig 2b. VIX and TS (correlation= 0.11)

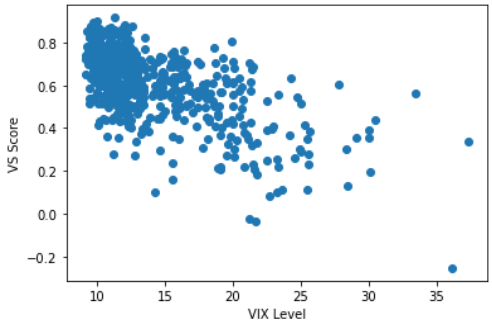

Fig 2c. VIX and VS (correlation= -0.62)

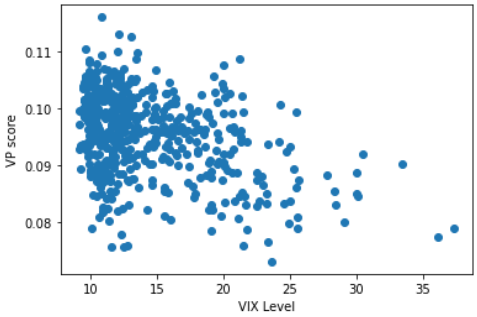

Fig 2d. VIX and VP (correlation= -0.39)

Figure 2. Correlation between different sentiment scores and VIX level.

Using the Pearson correlation coefficient, shown in Figure 2, the correlation between the scaled VIX and TP, TS, and VS are negative. These results align with the literature (Xu and Cohen, 2018; Yang et al., 2018). The first three Pearson correlation coefficients express the relation between higher VIX levels and investor sentiment scores. As for the subjective score (TS) shown in Figure 2b, we would expect it not necessary to be correlated to the VIX level since it does not reflect positive or negative sentiment scores.

3.2.2 Calculating VIX level

We used the VIX series' raw data and capture "trend," which is calculated by the moving average of the last three trading days. The simple moving average (MA) of the VIX level on the day t is given by:

$$MA_t = \frac{VIX_{t-2} + VIX_{t-1} + VIX_t}{3}$$

Then, for the LSTM model, we subtract the trend from the raw data level to find the "noise", which we define as $x_t = VIX_t - MA_t$.

3.2.3. Input vector for training and trading sets for LSTM model

For any given day t, we define a vector L_t of size 5, including input sets of the last previous trading day. Each column includes "noise" of VIX index level and sentiment scores TP, TS, VS, VP.

$$L_t = \begin{bmatrix} x_t \\ TP_t \\ TS_t \\ VS_t \\ VP_t \end{bmatrix} \tag{1}$$

This input vector is used to forecast the VIX index level "noise" $\overline{x_{t+1}}$ for the day $t+1$ for the LSTM algorithm.

3.2.4. Input value for training and trading sets for ARIMA model

For any given day t, we define the input value MA_t of trading day t of VIX level input to forecast the VIX trend $\overline{MA_{t+1}}$ for day $t+1$.

3.2.5. Splitting the input vector into in-sample and out-of-sample sets

Following Krauss et al. (2017) and Fischer and Krauss (2018), we define an "in-sample" training period as a set, consisting of a training period of 718 days (approximately 70% of the dataset), which is equivalent to nearly three years, and an "out-of-sample" trading period of the succeeding 309 days (30% of the dataset). Therefore, for the LSTM model, the size of the training set is (5,718), meaning 718 L_t vectors, each of size 5, and the trading set size is (5,309), meaning 309 L_t vectors

of size 5. For the ARIMA model, the training set size is (1,718), and the trading set size is (1,309). Figure 3 shows the VIX index level over time. The blue line describes the training set (70% of the dataset), while the orange line describes the trading set (30% of the dataset).

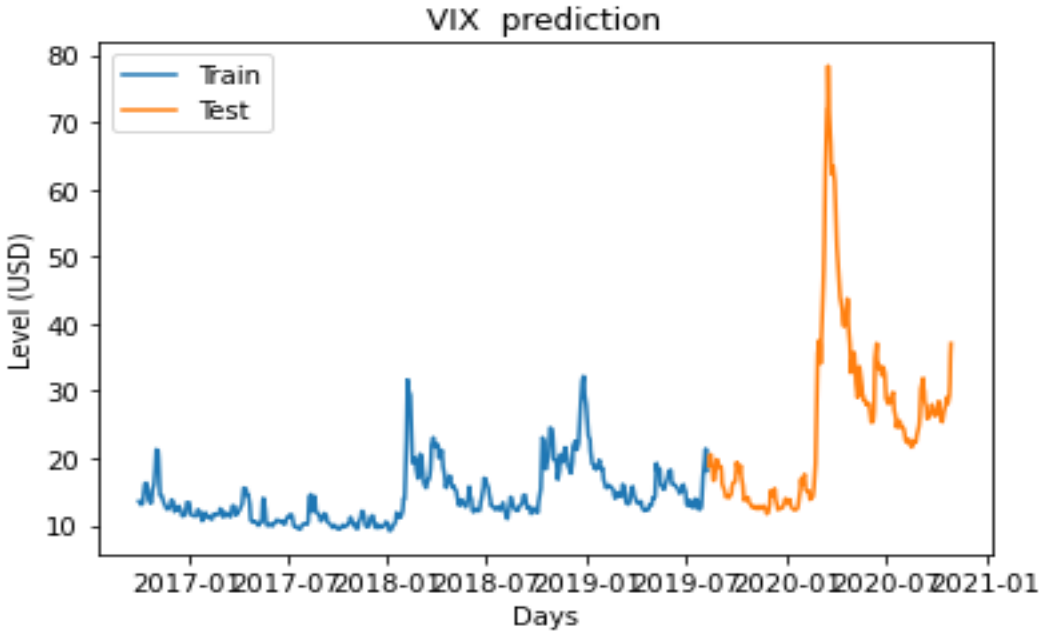

Figure 3. VIX Index and the allocation for the training set and test set

Figure 3 shows that both the training set and the trade set consist of significant VIX index spikes. The COVID-19 outbreak during March 2020 associated with the spike in VIX index level is included within the trading set.

3.3 In-Sample analysis

For the in-sample analysis, the input data is used to find the best parameters that calibrate the model and best fit them in terms of the lowest error between theoretical VIX index level and actual VIX index level. In this stage, we separate the in-sample analysis for the LSTM model and the in-sample analysis for the ARIMA model. First, we detail the in-sample process for both modes (i.e., LSTM model and ARIMA model), features, and architecture separately. Then, we introduce the hybrid LSTM-ARIMA model.

3.3.1 LSTM model

LSTM model has been introduced by Hochreiter and Schmidhuber (1997) and further refined in the following years by Gers et al. (2000) and Graves and Schmidhuber (2005), to name a few. LSTM networks are specifically designed to learn long-term dependencies and overcome the previously inherent problems of RNNs, i.e., vanishing and exploding gradients (Sak et al., 2014).

LSTM networks are composed of an input layer, one or more hidden layers, and an output layer. The number of neurons in the input layer is equal to the number of explanatory variables (input vector), which in our model is L_t. The number of neurons in the output layer reflects the output space, which in our model is one neuron represent the VIX index level "noise" $\overline{x_{t+1}}$.

The hidden layers in LSTM networks consist of memory cells. Each of the memory cells has three gates maintaining and adjusted its cell state: an input gate, an output gate, and a forget gate. At every time step *t* each of the three gates acts as filters of the information obtained from the previous layer. For more details, the reader is referred to Fischer and Krauss (2018). Every neural network, such as the LSTM model, has a loss function and an optimizer function. The loss function is the error between the actual output and the predicted output. For accurate predictions, one needs to minimize the calculated error. In a neural network, minimizing loss function is carried using backpropagation. The current error is typically propagated backward to a previous layer, where it is used to modify the weights and biases so that the error is minimized. The weights are adjusted using a function called Optimization Function.

We apply two advanced methods for the LSTM model training; each of them uses *Keras* (an open-source neural-network library written in Python). First, we make use of Nesterov accelerated adaptive moment (Nadam) as an optimizer. Second, we use absolute mean error as the loss function in all the experiments, as the absolute mean error produces minimum loss during the training. The specified topology of our trained LSTM network is specified below (see Figure 4):

- Input layer L_t.
- Two LSTM hidden layers, each with h = 30 hidden neurons and a dropout value of 0.2.
- Output layer (dense layer) with one neuron representing the forecast for day t+1 using the linear activation function.

Following Gal and Ghahramani (2015), we apply dropout regularization within each of the two hidden layers. Because of this, 20% of the input units are randomly dropped at each update iteration during training time, both at the input gates and the recurrent connections, resulting in reduced risk of overfitting and better generalization. The training samples were split into two sets: one training set and one validation set. We kept about 25% of the in-sample dataset as a validation set (these samples are assigned randomly to either training or validation set). The first set is used to train the network and iteratively adjust its parameters to minimize the loss function. The second set of the network predicts the unseen samples from the validation samples and try to forecast the VIX index level and validate the selected parameters.

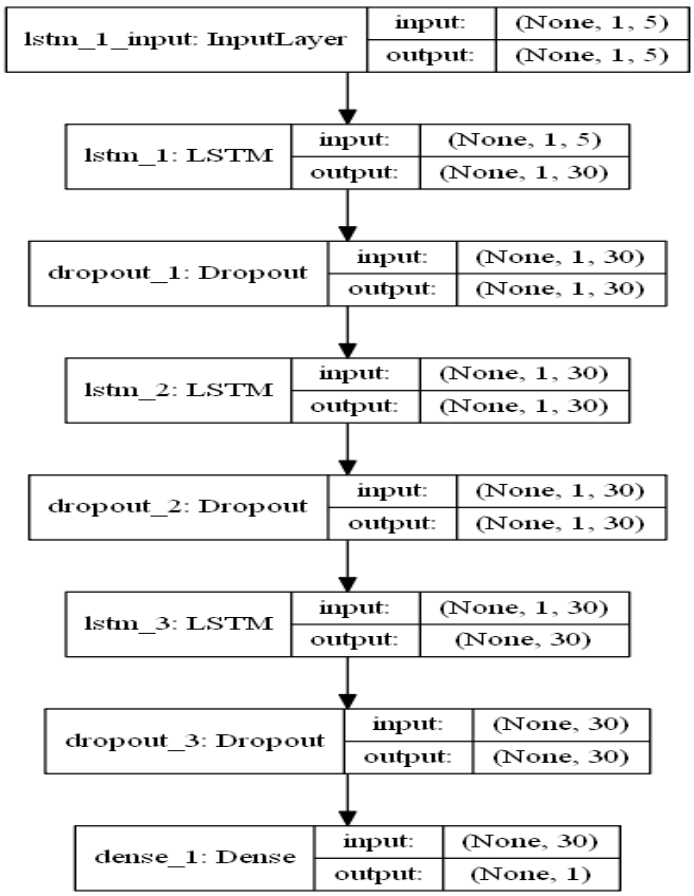

Figure 4. LSTM model topology, inputs, and outputs for each layer

Figure 4 shows our model topology, inputs, and outputs for obtaining the optimal parameter weights for the model. There is a total of 18991 weights parameters estimated for calculating the optimal VIX for *t+1*.

3.3.2 ARIMA

The ARIMA model (Koreisha and Fang, 1999; Cline and Brockwell, 1985) is a generalization of an Auto-Regressive Moving Average (ARMA) model, with an integrated component as a measure of how many non-seasonal differences are needed to achieve stationarity. Both models use time-series data to better understand the data or forecast future points in the series, based on a combination of two polynomials, one for the autoregressive part and the other for the moving average part.

Following Musa and Joshua (2020), we used ARIMA (1,1,1) to forecast the next day's value. The first element is the order (number of time lags) associated with the autoregressive model. The second element is the number of differencing (subtract the previous value from the current value) required to make the time series stationary. The third element is the order associated with the moving-average model. The Input value for our model is x_t

3.4 Hybrid LSTM-ARIMA model

As the ARIMA approach individually have a single forecast for the VIX index level in *t+1 and LSTM model approach individually have a single forecast* for "noise" from the trend, hybridizing the ARIMA and LSTM forecast for t+1 by adding them will provide a forecast to VIX index levels: $VIX_{t+1} = \overline{x_{t+1}} + \overline{MA_{t+1}}$ By doing so, we capture both the trend and the specific error derived from the four sentiment scores and the interconnections in the VIX index itself for recent days.

4. OUT-OF-SAMPLE EMPIRICAL RESULTS

For all models, we've used RMSE, MAPE, and MAE to measure the efficiency of the suggested method in forecasting the actual VIX index level in *t+1*. Low RMSE, MAPE, and MAE scores imply better forecasting.

4.1 Out-of-Sample performance analysis – gap analysis

Our results include three steps. First, we analyze RMSE and MAE for the Hybrid LSTM-ARIMA approach. We then analyze each model's profitability, separately, for

both models, and last, we perform a robustness test on the Hybrid LSTM-ARIMA results.

We have used both ARIMA and LSTM and LSTM without sentiment analysis (only one historical VIX price as input feature). Table 2 shows the empirical results. We used RMSE and MAE, following Zhu and Lian (2012) and others.

Table 2. Out of sample results in 6 different models

Model	RMSE	MAE
Hybrid LSTM-ARIMA with sentiment	3.00681	1.7069
Hybrid LSTM-ARIMA without sentiment	3.01327	1.72781
ARIMA	3.18904	1.73033
LSTM with sentiment	10.8492	5.44829
LSTM without sentiment	10.997	5.71967

Table 2 shows that Hybrid LSTM-ARIMA with sentiment has the lowest RMSE score and MAE score. Very nearby results are achieved by the ARIMA model results, while the LSTM models obtain inferior results.

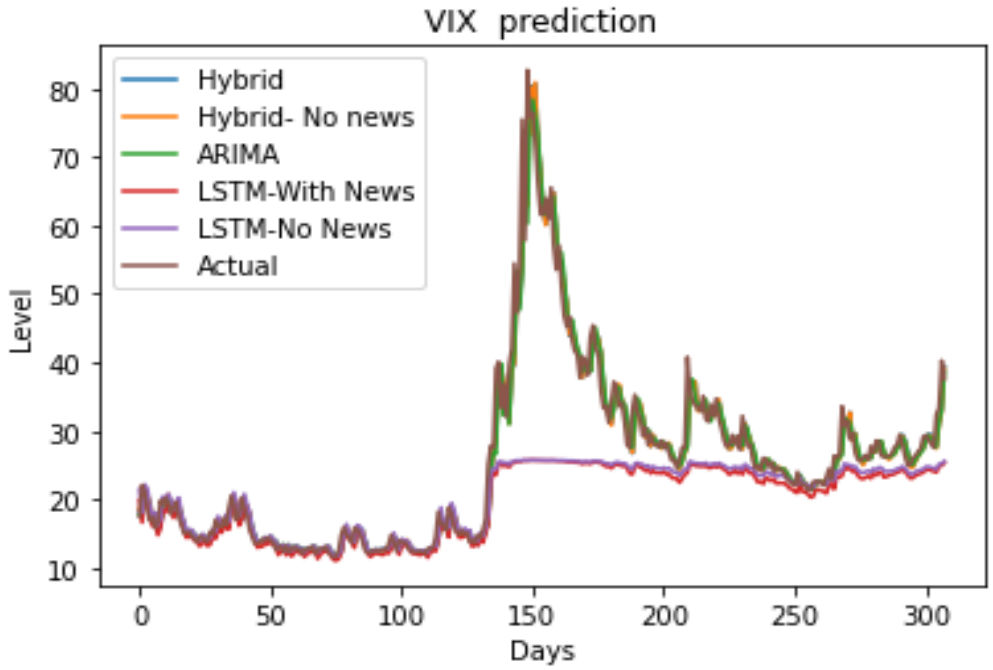

Figure 5. VIX Index level forecast for hybrid models with (and without) sentiment compared to VIX index levels

Figure 5 shows that except for the LSTM models solely, the VIX index level forecast is very close to actual values, even during COVID-19 trading day, which had extremely high spikes.

Interestingly, before COVID-19, all models show high accuracy to the actual VIX levels. However, during COVID-19 trading days, the LSTM component didn't capture the higher volatility. We conclude that the investors' sentiment increases the LSTM model accuracy and the ARIMA model helps capture the short time previous results more accurate in times of higher fluctuation.

4.1 Out-of-Sample performance analysis - Trading strategy

For trading strategy, the forecast at *t+1*:

- According to the model, if $VIX_{Model,t+1} > VIX_t$ buy the VIX index and the profit (loss) on day *t*+1 will be $\pi_{t+1} = -VIX_{Actual,t} + VIX_{Actual,t+1}$.
- According to the model, if $VIX_{Model,t+1} < VIX_t$ sell the VIX index and the profit (loss) on day *t*+1 will be $\pi_{t+1} = VIX_{Actual,t} - VIX_{Actual,t+1}$.
- According to the model, if $VIX_{Model,t+1} = VIX_t$ do not trade.

We present in Table 3 the empirical trading strategy results before any transaction costs.

Table 3. Trading strategy results for each model

Model	Total profit (in $)	% of Buy Transactions
Hybrid LSTM-ARIMA with sentiment	151.17	44%
Hybrid LSTM-ARIMA without sentiment	117.57	41%
ARIMA	85.43	56%
LSTM with sentiment	4.65	32%
LSTM without sentiment	-20.07	0%

Table 3 shows that Hybrid LSTM-ARIMA sentiment obtains the highest total profit (in $). The cumulative profit is shown in Figure 6.

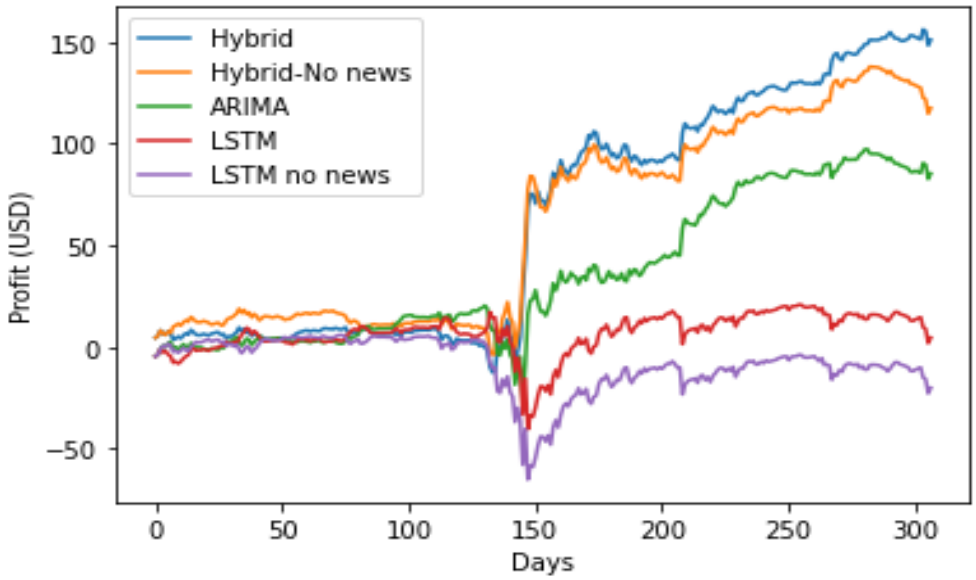

Fig 6. VIX cumulative profit (in $) for all tested models

Figure 6 shows that the hybrid LSTM with sentiment obtains a significantly higher profit of 29% than the hybrid LSTM without sentiment, mainly during COVID-19 volatile time. The inclusion of the COVID-19 period is interesting because it provides insight of the power of the model even in the face of unexpected shocks.

4.2. Robustness test results

Several methods to compute robustness quantification of several neural networks for Out-of-Sample data were presented (Deng et al., 2016; Ko et al., 2019). Yet, the robustness quantification of LSTM models for out-of-sample data remains an open problem because of the complexity of its architecture. To quantify robustness, we made the following robustness random shuffling test on the Hybrid LSTM-ARIMA with the sentiment on the Out-of-Sample data. By doing so, we address the main vulnerability of LSTM models: *The adversarial attack-based approach*, which means researchers design strong adversarial attack algorithms to attack deep neural networks. Robustness is measured by the distortion between successful adversarial examples and the original ones. The following three Hybrid LSTM-ARIMA models are baseline Hybrid LSTM-ARIMA model with the sentiment, baseline Hybrid LSTM-ARIMA model with VIX data randomly shuffle, baseline Hybrid LSTM-ARIMA model with both VIX and sentiment score values data randomly shuffled. By shuffling the data, we can assure that the results are not arbitrary. Table 4 shows the robustness test results for the three models.

Table 4. Robustness test for out-of-sample hybrid LSTM-ARIMA baseline model

	Baseline Model (No data shuffle)	**Baseline model with shuffle VIX values**	**Baseline model with shuffle VIX and sentiment score values**
RMSE	3.00681	3.15618	3.19327
MAE	1.7069	1.9134	1.92676
Profit (USD)	151.17	116.25	108.05
% of Buy Transactions	44%	48%	45%

Table 4 shows that the baseline model (no data shuffle) has robust results in all parameters (RMSE, MAE, Profit) than the two shuffled data models. RMSE and MAE are the lowest, indicating that the results are robust. The profit ($) of the baseline model is more significant than the two shuffled data models.

5. CONCLUSION

This paper introduces a new forecasting model of VIX index returns for the next day based on both LSTM and ARIMA models. We developed the hybrid LSTM-ARIMA model, which considers investors' sentiment scores. To the best of our knowledge, this study is the first to propose and implement a hybrid LSTM-ARIMA model and to incorporate investors' sentiment analysis in the model.

The sentiment scores are empirically evaluated based on commonly used daily article text major economic sites. The forecasts of next day VIX index level based on out-of-sample for 2019-2020 end of day data present robust results compared to models without sentiment parameters. We found that hybrid LSTM-ARIMA with sentiment obtains the lowest RMSE, while the LSTM model (with or without sentiment) obtains inferior results, mainly in the COVID-19 period. We relate these outcomes to the fact that the LSTM model has a positive bias towards actual prices.

The model introduced in this paper has significant advantages and implications for trading the VIX index. The model is unique by combining LSTM-ARIMA for VIX forecasting and adding a sentiment analysis methodology to improve the empirical results. The model also achieves greater proximity to the actual VIX index levels and higher profits than each of the comparative methods under consideration in this study.

Importantly, our study's main conclusion is that sentiment analysis improves forecasting compared to hybrid LSTM-ARIMA without sentiment.

Some of the main limitations of this work can be considered for future research. The dataset included only investors' sentiment and VIX index levels as the input vector. We did not consider parameters such as trading volumes in the VIX index or in the S&P500 index to increase proximity or profit. Our analysis was made on end-of-day data, but for future research it would be scientifically interesting to extend the work and perform the analysis of intraday data.

ACKNOWLEDGMENTS: We would like to thank participants of Centre for Global Finance Reading Group, School of Finance and Management, SOAS university of London Reading Group who provided insight and expertise that greatly assisted the research.

DATA AVAILABILITY STATEMENT: The data that support the findings of this study are available from corresponding author

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation (16) (pp. 265-283).

Ahoniemi, K. (2006). Modeling and forecasting implied volatility-an econometric analysis of the VIX index. Helsinki: Helsinki Center of Economic Research.

Chollet, F. (2015). Keras.

Cline, D. B., & Brockwell, P. J. (1985). Linear prediction of ARMA processes with infinite variance. *Stochastic Processes and their Applications*, 19(2), 281-296.

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3), 653-664.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.

Froot, K. A., Scharfstein, D. S., & Stein, J. C. (1992). Herd on the street: Informational inefficiencies in a market with short-term speculation. *The Journal of Finance*, 47(4), 1461-1484.

Gal, Y., & Ghahramani, Z. (2015). A theoretically grounded application of dropout in recurrent neural networks. *arXiv preprint arXiv:1512.05287*.

GERS, F. A., SCHMIDHUBER, J., & CUMMINS, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10), 2451-2471.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), 602-610.

Guo, S., & Liu, Q. (2020). Efficient Out-of-Sample Pricing of VIX Futures. *The Journal of Derivatives*, 27(3), 126-139.

Hao, J., & Zhang, J. E. (2013). GARCH option pricing models, the CBOE VIX, and variance risk premium. *Journal of Financial Econometrics*, 11(3), 556-580.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2), 327-343.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Jin, Z., Yang, Y., & Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 1-17.

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, 11(2), 2664-2675.

Koreisha, S. G., & Fang, Y. (1999). The impact of measurement errors on ARMA prediction. *Journal of Forecasting*, 18(2), 95-109.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38-48.

Lalancette, S., & Simonato, J. G. (2017). The role of the conditional skewness and kurtosis in VIX index valuation. *European Financial Management*, 23(2), 325-354.

LeBaron, B. (2000). Agent-based computational finance: Suggested readings and early research. *Journal of Economic Dynamics and Control*, 24(5-7), 679-702.

Loria, S. (2018). Textblob Documentation. Release 0.15, 2.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56).

Musa, Y., & Joshua, S. (2020). Analysis of ARIMA-Artificial Neural Network Hybrid Model in Forecasting of Stock Market Returns. *Asian Journal of Probability and Statistics*, 42-53.

Psaradellis, I., & Sermpinis, G. (2016). Modeling and trading the US implied volatility indices. Evidence from the VIX, VXN and VXD indices. *International Journal of Forecasting*, 32(4), 1268-1283.

Qi, M., & Wu, Y. (2006). Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. *Journal of Money, Credit and Banking*, 2135-2158.

Qiao, G., Yang, J., & Li, W. (2020). VIX forecasting based on GARCH-type model with observable dynamic jumps: A new perspective. *The North American Journal of Economics and Finance*, 53, 101186.

Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Fifteenth annual conference of the international speech communication association.*

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1394-1401).

Taylor, N. (2019). Forecasting returns in the VIX futures market. *International Journal of Forecasting*, 35(4), 1193-1210.

Urologin, S. (2018). Sentiment Analysis, Visualization and Classification of Summarized News Articles: A Novel Approach. *IJACSA0 International Journal of Advanced Computer Science and Applications*, 9. (8).

Walt, S. V. D., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22-30.

Wang, T., Shen, Y., Jiang, Y., & Huang, Z. (2017). Pricing the CBOE VIX futures with the Heston–Nandi GARCH model. *Journal of Futures Markets*, 37(7), 641-659.

Xu, Y., Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational . 1,* 1970-1979.

Yang, X., & Wang, P. (2018). VIX futures pricing with conditional skewness. *Journal of Futures Markets*, 38(9), 1126-1151.

Yang, L., Zhang, Z., Xiong, S., Wei, L., Ng, J., Xu, L., & Dong, R. (2018). Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 441-445).

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.

Zhu, S. P., & Lian, G. H. (2012). An analytical formula for VIX futures and its applications. *Journal of Futures Markets*, 32(2), 166-190.