Problem of Entrepreneurship, trade and exchange activity

UDC 336.64

Symonenko Olena
Candidate of Economic Sciences, Associate Professor
National University of Life and Environmental Sciences of Ukraine
ORCID: 0000-0002-2459-4187

Symonenko Dmytro
Master of Science of the
University of Vienna

FORECASTING STOCK MARKET FOOD INDEX USING COMMODITY FUTURES QUOTES

Summary. *This article is devoted to testing whether commodity futures can act as tools to predict price fluctuations of commodity indices. We decided to choose WIG-food index and leading commodity futures since they are interconnected. Companies included in our index use commodities in their operations. While we managed to find comparable articles regarding financial market forecasts using different indices, we did not find any opportunities to predict niche indices like WIG-food in our example. Therefore, our article may act as clarification of whether it is possible to use machine learning algorithms combined with basic regression to predict commodity indices price fluctuations.*

The prediction of stock price movements has long been an intriguing topic of financial research. Particularly prominent ways of predicting a stock price trend include using stocks past performance or using sentiment analysis. However, a certain level of criticism must be applied to these methods. Using past performance creates a certain degree of isolation that leaves out important information carried out by other entities and makes the prediction result vulnerable to local perturbations.

Before, we tested whether the relationship between these variables exists to implement machine learning algorithms in price forecasting. Our study consists of Introduction, where were familiarize the reader with the topic; a Methodology, where we describe data collection principles and reasoning for conducting the research, as well as describing machine learning algorithms behind our analysis; Results, where we present the outcomes of our models and a Conclusion. We proposed the use of data collected from different global financial markets with machine learning algorithms to predict stock index movements. In this project, we have attempted to forecast stock price trends using machine learning techniques LSTM. Before we ran basic regression using Keras to derive the reasonability of applying the LSTM model. Regression results showed that commodities included in the model were far from perfection and we managed to achieve 48% accuracy of regression predictive power using commodity futures as features. Therefore, we excluded them from the LSTM model since they turned out to be not credible variables to apply in the machine learning algorithm.

***Key words:** index, commodity futures, correlation, regression, epoch, accuracy, mean absolute errors, machine learning, recurrent neural network (RNN), long short term memory (LSTM).*

Statement of the problem. Investors are looking for tools to predict stock market fluctuations. With rising awareness of financial markets, it is important to find credible indicators and a set of approaches to make efficient financial decisions. Therefore, we decided to devote our analysis to predicting stock index price fluctuations based on futures quotes. We believe commodity futures can act as a good predictor of commodity indices since futures describe how speculators or investors see the market in the upcoming days. We decided to choose WIG-food index and leading commodity futures since they are interconnected. Companies included in our index use commodities in their operations. While we managed to find comparable articles regarding financial market forecasts using different indices, we did not find any opportunities to predict niche indices like WIG-food in our example. Therefore, our article may act as clarification of whether it is possible to use machine learning algorithms combined with basic regression to predict commodity indices price fluctuations.

Analysis of recent researches and publications. The majority of the articles we looked for are based on leading indices like S&P500 or NASDAQ, while niche indices like WIG are staying unnoticed. Therefore, we believe our work will add value to previous studies from: Cho K. et al [1], Shunrong J. et al [4] and Wu J. et al [8]. We collect our data from popular financial websites like YahooFinance and Investing.com using API. All calculations and data visualizations were performed in Python. In our analysis we run basic Keras regression to verify reasonability for applying machine learning algortihms. More details are included in “Methodology” section.

Formulation purposes of article (problem). The prediction of stock price movements has long been an intriguing topic of financial research. Particularly prominent ways of predicting a stock price trend include using stocks past performance or using sentiment analysis. However, a certain level of criticism must be applied to these methods. Using past performance creates a certain degree of isolation that leaves out important information carried out by other entities and makes the prediction result vulnerable to local perturbations. As for sentiment analysis, it may fail due to misleading and biased information processing (Wu et al., 2021). As a result, after consulting with the paper of Shen and Jiang (2012), it was decided to test and analyse the possibility of using machine learning techniques together with commodity futures to predict stock market indices fluctuations. Our research is based on the belief that stock market indices include companies whose operations involve commodities, such as grain or egg producers for example.

The report is structured in the following way. First, the setting of the study is constructed, and the data collected is presented. Afterward, we test whether basic regression can be applied to our variables to verify whether it is reasonable to apply machine learning algorithms. Then, we implement long short-term memory (LSTM) network. Finally, we present our findings and specify conclusions obtained from the model.

The main material.

Reasoning & Data Collection

We are using the Polish WIG-food index as our forecasting target. WIG-food is a sub-sector index, and its portfolio includes WIG constituents belonging to the "food and drinks" sector. Weightings in the index are the same as in the WIG index portfolio. WIG-food index base date is December 31, 1998. The initial value was 1279.56 points. It is an income-based index and thus when it is calculated it accounts for both prices of underlying shares and dividend and pre-emptive rights" income [7]. Companies included in this index can be found in Table 1.

Table 1

List of companies included in the WIG-food index

Company	ISIN	Number of shares in the index	Number of shares in the index (PLN)	Share in index (%)
KERNEL	LU0327357389	52,057,000	1,093,197,000	48.816
WAWEL	PLWAWEL00013	509,000	220,906,000	9.864
AMBRA	PLAMBRA00013	9,800,000	194,236,000	8.673
ASTARTA	NL0000686509	7,522,000	180,528,000	8.061
IMCOMPANY	LU0607203980	8,468,000	141,415,600	6.315
TARCZYNSKI	PLTRCZN00016	2,835,000	111,699,000	4.988
OVOSTAR	NL0009805613	1,378,000	60,907,600	2.720
PEPEES	PLPEPES00018	25,009,000	37,263,410	1.664
GOBARTO	PLDUDA000016	3,772,000	28,290,000	1.263
PAMAPOL	PLPMPOL00031	6,502,000	27,698,520	1.237

Company	ISIN	Number of shares in the index	Number of shares in the index (PLN)	Share in index (%)
MAKARONPL	PLMKRNP00015	2,974,000	23,078,240	1.031
KRVITAMIN	PLKRVTM00010	1,729,000	21,785,400	0.973
OTMUCHOW	PLZPCOT00018	8,783,000	21,430,520	0.957
SEKO	PLSEKO000014	2,494,000	15,712,200	0.702
ATLANTAPL	PLATLPL00018	1,765,000	15,179,000	0.678
KSGAGRO	LU0611262873	6,282,000	15,139,620	0.676
AGROTON	CY0101062111	3,163,000	13,031,560	0.582
HELIO	PLHELIO00014	934,000	10,274,000	0.459
MILKILAND	NL0009508712	6,694,000	6,091,540	0.272
MBWS	FR0000060873	142,000	1,562,000	0.070

Source: GPW Benchmark [7]

Since companies included in the WIG-food index operate in the "food and drinks" industry, we therefore will test whether commodity futures can act as a forecasting tool to predict index fluctuations. We are using the following commodity futures as predictors of WIG-food index quotes: ZC (corn), ZO (oat), KE (wheat), ZR (rough rice), ZS (soya beans), GF (feeder cattle), LE (live cattle), HE (lean hogs), CC (cocoa), SB (sugar). Futures are derivative financial contracts that obligate parties to buy or sell an asset at a predetermined future date and price, which is in general used for hedging or speculating purposes.

We collect data used in our model from Yahoo Finance [9] and Investing.com [6] using API tools for the corresponding web pages to scrape data fast. We conduct our analysis using Python and topic-related libraries. We collected data from 03-10-2011 to 23-02-2022. We chose 03-10-2011 specifically since all commodity futures were quoted at that specific date since before some of them were not available for trading. We also eliminate trading days after 23-

02-2022 from our model due to the Russia-Ukraine conflict which led to significant price fluctuations of agricultural commodities and caused considerable damage to companies included in the WIG-food index, since most of them are from Ukraine. Moreover, almost 49% of WIG-food index price depends on Kernel, which is the world’s leading and Ukraine’s largest producer and exporter of sunflower oil, and a major supplier of agricultural goods, meaning Russia-Ukraine conflict had a significant impact on price fluctuations since Kernel’s production is located in Ukraine and was damaged as a result of the conflict [2].

Machine Learning Algorithm: Long Short Term Memory Network (LSTM). The machine learning algorithm introduced in our project is LSTM. LSTM is a type of Recurrent Neural Networks (RNNs) used for dealing with long-term dependencies. Basic RNNs use the reasoning regarding previous events to inform events in the future, mainly predicting their outcome. RNNs are used for multiple purposes, starting from speech recognition and language modeling to forecasting stock market movements, which is the aim of our paper.

RNNs algorithms are based on loops which leads to information persistence. In the below diagram (Fig. 1), a piece of Neural network A, loops into the input Xt and returns a value ht.

RNN Algorithm Description

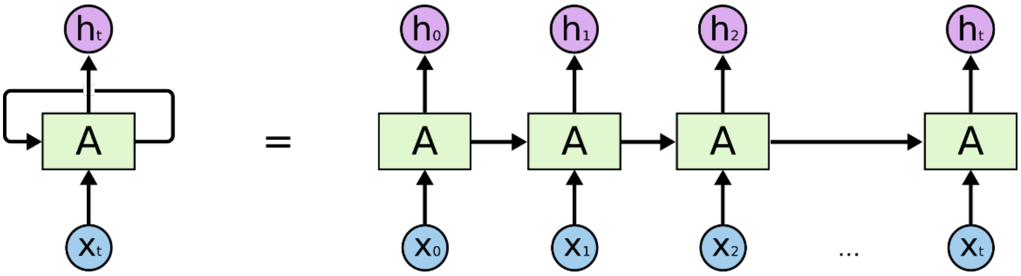


Fig. 1. An unrolled recurrent neural network

Source: colah’s blog [3]

A loop allows information to be passed from one step of the network to the next. This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists. They are the natural architecture of neural network to use for such data. (Olah C., 2015). Every additional input in the RNN model improves the accuracy of the next output, thus, training dataset for such models must be substantial to return consistent predictions. The biggest strength of RNN is its ability to connect previous information and produce current and future values. However, for big databases basic RNNs are not persistent: at some point we may need information from the beginning of the data to produce desired output, but with growing gap between between relevant information and the point where it is needed, the RNN will be unable to keep the connection due to its short-term memory. To solve this problem, we are using a Recurrent Neural Network capable of dealing with long-term dependencies - Long Short Term Memory Network or LSTM [Figure 2]. LSTMs can remember the information for long periods while steadily increasing its learning curve.

LSTM Algorithm Description

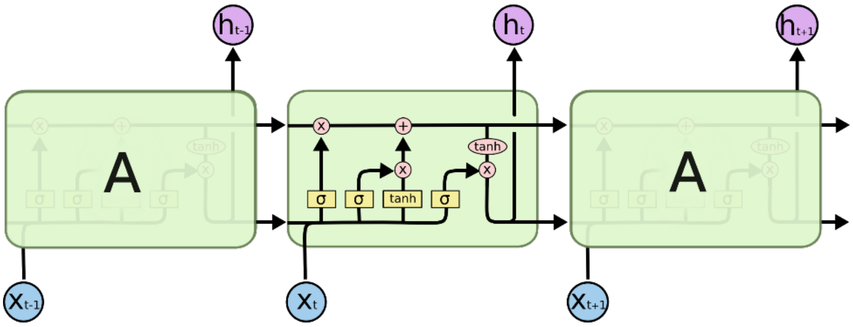


Fig. 1. The repeating module in an LSTM contains four interacting layers.

Source: colah's blog [3]

ML Algorithms Notation

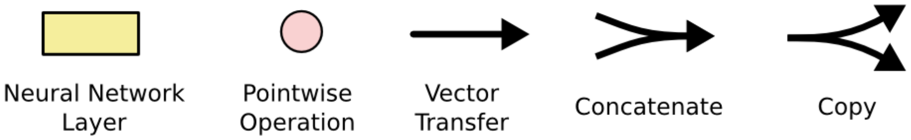


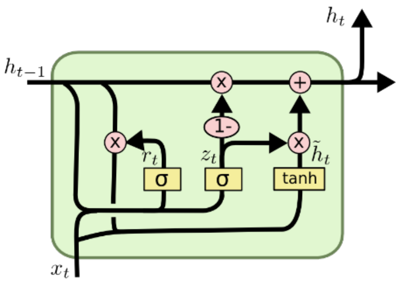
Fig. 3. Notation used for ML process description

Source: colah's blog [3]

All RNNs are chains of repeating modules of neural networks, which transmit information from one to another making all the necessary computations inside every module. In basic RNNs modules include simple computations, however, modules in LSTMs are more complex and have multiple layers (yellow rectangles in the diagram [Figure 3]) which interact with each other. Data produced by these interactions is then transferred to the next module by chains and the procedure repeats again but with more accurate results.

In our model, we have used one of the most popular versions of LSTM - Gated Recurrent Unit (GRU), introduced by Cho, et al. (2014). This model combines the forget gates and input gates. Forget gates look at the values ht−1 and xt and makes a decision whether to keep this data or get rid of this. Input gates follows the same procedure but the decision to be made is whether to update the value based on the information from the previous module or not. GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets. As we are using low-frequent (daily) data for our project, we have chosen this specific neural network.

Gate Recurrent Unit (GRU) Description



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Fig. 4. Computations within GRU gate

Source: colah’s blog [3]

We have used LSTM model to predict market movements (which is the WIG-food index) using the quotes from leading commodity futures (corn, oat, wheat, rice, soya, cattle, hogs, cocoa, and sugar). Additionally, we used the 50 moving averages to predict our target more precisely. In total for LSTM model, we used 8 features to predict open, low, high, and close daily quotes of the WIG-food index. Before, we explored the correlation between every feature and target data to be sure that there are dependencies in our future model. Next, we used Keras Regressor with epochs=100 to run basic regression on our dataset. Further, data was converted using MinMaxScaler and 30% of it was used to train the model. We added four layers to our LSTM model and used Adam Optimizer to deal with the loss function (mean squared error). The results of the model will be discussed in the corresponding section.

Insights from this study and perspectives for further research in this direction. Basic regression ran by Keras Regressor showed positive results with few outliers and most of the points lying on the regression line, which proves that data is consistent to run the LSTM (Fig. 4, 5, 6, 7). Basic regression showed the 47.6% of accuracy in the 99th epoch with a value of the mean absolute error of 115.6 in the last epoch (Fig. 8, 9). The accuracy rate was increasing steadily after every epoch while the mean absolute error was falling until reaching the minimum value of 123.94 in the 98th epoch.

Basic regression showed worse results than we expected since we managed to achive only 49% prediction accuracy at maximum. Moreover, the mean abosulte error seems pretty high for this kind of analysis. Possible explanation can be that some commodity futures are not strongly correlated with WIG-food index due to the small fraction of operations in the companies included in the index. While commodity futures have positive correlation with WIG-food index, it is still not strong leading to the conclusion that our model may be overfit with prediction factors and some of them should be eliminated.

Therefore, we run our LSTM model excluding features which include commodity futures prices due to their poor correlation with WIG-food index. LSTM model was run to predict high, low, open and close values for WIG-food index based on moving average and trading volume features. We received the chart below which proves that LSTM model is able to efficiently identify the stock index patterns and predict values close to the actual ones. The explained variance score of LSTM model is 92% while R2 is 79%. However, predicted values lie far from perfection, thus, our model is subject to limitations which will be discussed in the end of this paper.

Regression Results

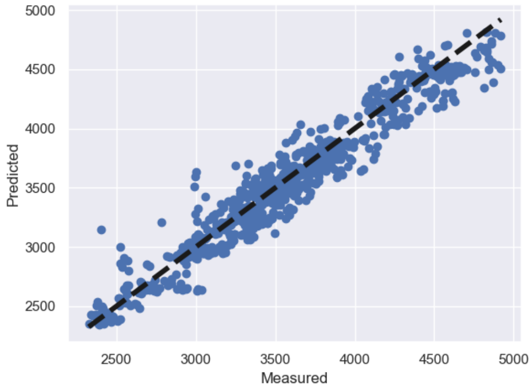


Fig. 4. Scattered Plot of Predicted/Measured High Price

Source: own calculations

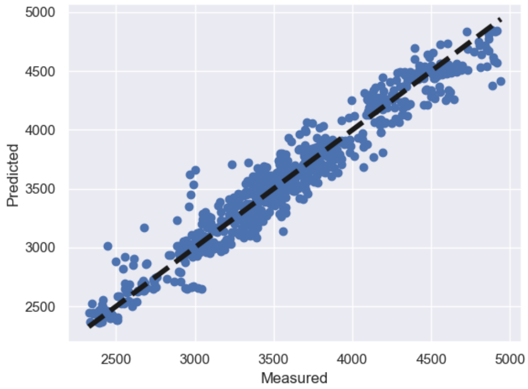


Fig. 5. Scattered Plot of Predicted/Measured Low Price

Source: own calculations

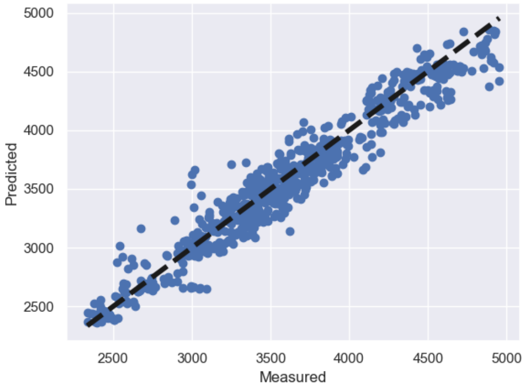


Fig. 6. Scattered Plot of Predicted/Measured Close Price

Source: own calculations

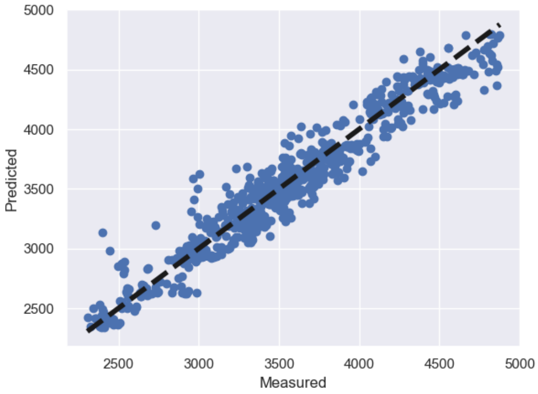


Fig. 7. Scattered Plot of Predicted/Measured Open Price

Source: own calculations

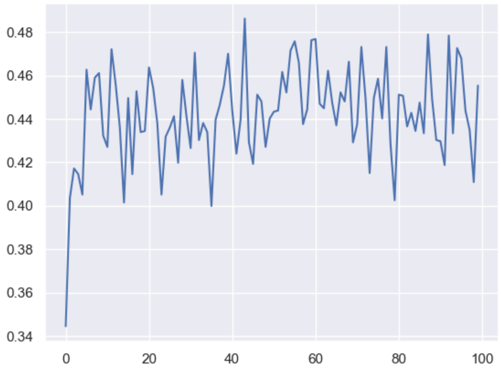


Fig. 8. Regression accuracy with rising epochs

Source: own calculations

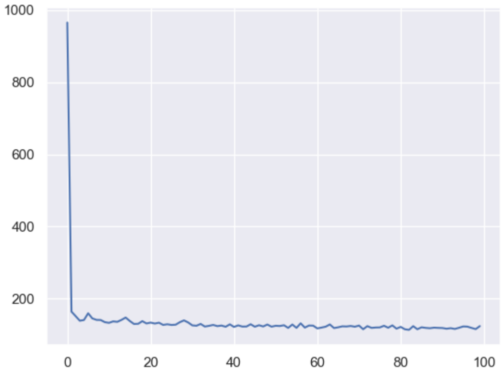


Fig. 9. Regression mean absolute error timeline

Source: own calculations

Predicted WIG-food Index Using LSTM Algorithm

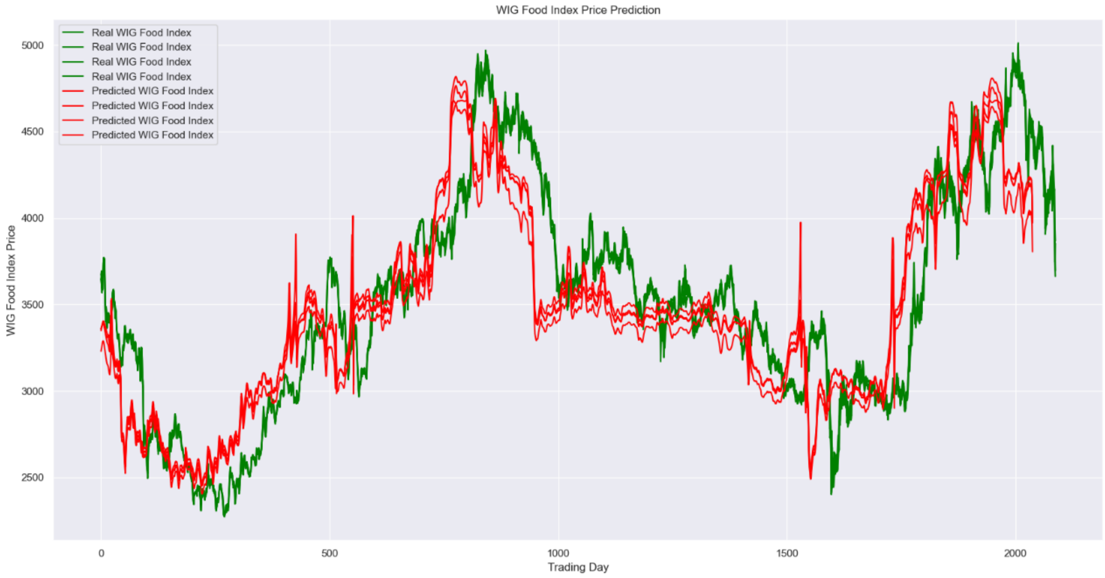


Fig. 10. LSTM Model WIG-food index prediction

Source: own calculations

Conclusions. In the project, we proposed the use of data collected from different global financial markets with machine learning algorithms to predict the stock index movements. In this project we have attempted to forecast stock price trends using machine learning techniques LSTM. Before we ran basic regression using Keras to derive the reasonability of applying the LSTM model. Regression results showed that commodities included in the model were far from perfection and we managed to achieve 48% accuracy of regression predictive power using commodity futures as features. Therefore, we excluded them from the LSTM model since they turned out to be not credible variables to apply in the machine learning algorithm. Our findings can be summarized into two aspects:

1. Commodity futures can act as predictors for stock market fluctuations, however, the WIG-food index turned out to be far from being a good variable to predict using commodity futures quotes which can be visible from regression results. Moreover, correlation analysis ran before preparing regression analysis

showed there is no positive strong correlation between futures and WIG-food index.

2. Basic regression results showed that it is reasonable to run the LSTM model on our data. The model showed positive performance with strong accuracy. The R2 for the model was decreasing steadily after each epoch which helped the model to identify future stock index patterns. However, we excluded commodity futures from the algorithm, which boosted the efficiency of our prediction.

Limitations. The project is, however, not prone to limitations:

1. While using the freely available data, it was not possible to find hourly data for the studied assets for a longer period than two years. It resulted in us using daily data which may not be enough to capture price fluctuations since there is a trading day when the market experiences high volatility or low trading volumes.

2. We excluded prices after the Russia-Ukraine conflict since they significantly increase price volatility, especially in the agricultural industry. Moreover, companies present in the index are mostly from Ukraine, therefore, future studies can be devoted to measuring the predictability in high-volatility environments.

3. Futures are not quoted daily while the index is, therefore, we replaced missing quotes for futures with values for previous days which may have resulted in losing important market movements which in turn impacts our model accuracy. Therefore, high-frequency data should be used in future research.

4. We ran our models on 100 epochs, which in turn boosted the accuracy of our model after each epoch. However, the accuracy reached its limit at the last epochs, so there was no reasoning to expand the number of epochs used.

5. WIG-food index turned out to be a poor choice for conducting this kind of analysis since the majority of index price depends on Kernel share price (48% of the overall index). Since Kernel is not dealing with most commodity futures in their operations, the effect of the relationship between these futures and the index is weakened which is supported by our correlation analysis. Therefore, our

recommendation will be to substitute the WIG-food index with the weighted-average commodity index (like S&P500 Select Industry Indices).

References

1. Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. doi: https://doi.org/10.3115/v1/d14-1179
2. KERNEL. (n.d.). *KERNEL*. KERNEL. November 7, 2022. URL: https://www.kernel.ua/
3. Olah C. Understanding LSTM Networks. *GitHub*. August 25, 2015. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
4. Shunrong S., Jiang H., Zhang T. Stock Market Forecasting Using Machine Learning Algorithms. *Department of Electrical Engineering Stanford University*. 2010.
5. *Understanding LSTM Networks - colah's blog*. (n.d.). URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
6. *Wig Food Index (SPOZ)*. Investing.com. November 17, 2022. URL: https://www.investing.com/indices/wig-food
7. *WIG-food index*. (n.d.). GPW Benchmark. November 10, 2022. URL: https://gpwbenchmark.pl/en-karta-indeksu?isin=PL9999999888
8. Wu J. L., Huang M. T., Yang C. S., Liu K. H. Sentiment analysis of stock markets using a novel dimensional valence–arousal approach. *Soft Computing*. 2021. 25(6). P. 4433–4450. doi: https://doi.org/10.1007/s00500-020-05454-x
9. *Yahoo is part of the Yahoo family of brands*. (n.d.). URL: https://finance.yahoo.com/