DOI: 10.46541/978-86-7233-406-7_248

27th International Scientific Conference
Strategic Management
and Decision Support Systems
in Strategic Management
SM2022

Subotica (Serbia), 20th May, 2022

Ákos Barta
Doctoral School of Business Management,
Hungarian University of Agriculture and Life Sciences,
Gödöllő, Hungary
mail@bartaakos.com

Márk Molnár
Faculty of Economic Sciences, John von Neumann University
Kecskemét, Hungary

Zsuzsanna Naárné Tóth
Institute of Economics, University of Agriculture and Life Sciences,
Gödöllő, Hungary

INVESTIGATION OF THE ONLINE PRESS AND COMMODITY EXCHANGE USING NEURAL NETWORKS

Abstract: The printed press greatly influences people's consumption attitudes, their stock market decisions, and their information on the decisions and internal problems of certain organizations and companies. In other words, the information obtained in this way may encourage them to give up a certain stock market position or to establish new ones, as they think they are familiar with the willingness and tendencies to buy on the market. Therefore, it is important to analyze how close this relationship is, and how effectively the content of articles published in the press can be linked to the prediction of short-term exchange rate or price developments, and to see whether a formal method can be applied to investment decisions. With the help of neural networks, we can search for relationships between large amounts of diversified data sets, and to set up a contingent forward-looking price and/or exchange rate forecasting model.

Keywords: neural network, forecast, oil price, web scraping, article analysis

1. INTRODUCTION

There is no uniform oil price in the world, as the quality of the extracted crude oil varies and oligopolistic markets set different prices. Nevertheless the direction and magnitude of changes in crude oil prices is very similar, as different prices move together.

Due to the dynamic nature of supply and demand, the value of each benchmark is constantly changing. In the long run, a marker sold at a premium for another index may suddenly become available at a discount. (Arshad, Rizvi, Haroon, Mehmood, & Gong, 2021)

Roughly a quarter of all crude oil contracts in the world are for WTI Crude, and the U.S.'s current global market dominance is unquestionable. Thus, the WTI price was used. Official historical data has been downloaded from the U.S. Energy Information Administration (EIA) database. (EIA - Cushing, OK WTI Spot Price FOB (Dollars per Barrel), 2022.)

Web scraping, or web data collection or extraction, is a method of data acquisition used to extract data from websites. A typical Web scraper software can access the World Wide Web directly using the Hypertext Transfer Protocol or a web browser. While data collection can be done manually by the user, this term usually refers to automated processes implemented by a stick or web robot. It is a form of copying in which specific data is collected and copied from the web, typically to a central local database or spreadsheet, for later retrieval or analysis.

Web scraping of a website involves retrieving and extracting relevant data from it. A query is a download of a page (which the browser performs when a user views a page). Therefore, web crawling is a major component of web scraping that retrieves pages for later processing. Withdrawal may occur after retrieval. The content of a page can be analyzed, searched, reformatted, data can be copied to a spreadsheet or uploaded to a database. Web scrapers usually take something off the site to use for other purposes. An example is copying names and phone numbers, or companies and their URLs or email addresses to a list (collecting contacts).

In the early 1980s, a scientist named John Hopfield revived research in the field of ANNs. An associative model for neural networks has been proposed, which describes the storage of information as taking place between connections between neurons. Hopfield suggested that data processing be accomplished by "turning on" or "turning off" certain neurons depending on external stimuli. (Hopfield, 1982) This concept helped solve the problems originally described by Minsky and Papert. The model did this by suggesting that individual neurons cooperate with those around them. In other words, what happens to a single neuron typically happens to the surrounding neurons. These neural associations provide the foundation for pattern recognition, associative memory, and error correction while providing sufficient processing capacity to store information from large neural networks. (Aiyer, Niranjan, & Fallside, 1990)

Artificial neural networks (ANNs) are data processing systems that are based on and inspired by neurological networks in the brain. The systems are primarily used for sample identification and processing and are able to gradually improve performance based on the analytical results of previous tasks. (Jain, Mohiuddin, & Mao, 1996)

After the publication of Hopfield's network model, research on artificial neural networks has increased greatly. (Rojas, 1996). This manifested in advancements in computing systems and state-of-the-art technology. (Rabunal, 2005)

One such development by C. V. Soumya and Muzameel Ahmed was the use of mathematical algorithms to identify human gestures. Their paper examines an innovative method used to recognize and describe expressive body movements. (Soumya & Ahmed, 2017) These bodily movements have been observed mainly in Mudra, a classical dance practiced by Indian culture. Their goal was to build a system that uses pattern recognition and image processing that can easily identify specific human gestures, which in turn provide a description of these exercises and their health benefits.

ANNs have become an essential tool for predicting and identifying microbial accumulation in different environments. In an article Keaton (2017) described a number of methods that were used to obtain data from microorganisms. Its aim was to use the information gathered to expand knowledge to facilitate the performance of microbial fuel cells in the process of converting chemical potential energy from waste streams into electrical energy. (Farhat, Psaltis, Prata, & Paek, 1985)

ANNs have recently been used to forecast stock market performance. Kamran Raza (2017) has developed a number of techniques based on four different versions of ANNs. The main purpose of developing these techniques was to create a forecasting model that would facilitate the work of listed investors. To find the best forecasting model, several techniques were compared separately. The results showed that the behavior of the stock market can be predicted with an accuracy of up to 77%. (Huang, Nakamori, & Wang, 2005)

2. METHOD

2.1. Oil price analysis

In connection with the change in the price of oil, our goal is to determine the rise and fall of prices, and within this, to accurately identify and forecast turning points. As in the present research, the study aims at developing a method for the predictability of these turning points.

The moving average convergence divergence (MACD) is one of the most reliable and commonly used momentum indicators.

The method of calculation is relatively simple: the quotient of two moving averages with different number of periods. Several variants are known, currently the 12-day exponential moving average (EMA) is divided by the 26-day exponential moving average (EMA).

The value of the indicator fluctuates around unity. In most cases, the MACD is calculated by taking into account the difference between the two moving averages, but in the present case and in the trading method, the percentage shift shows the actual market movements.

In the course of the research, we ran several analyzes for Signal values, but this did not yield sufficiently sensitive results, so this was discarded. Subsequent research will focus on this and will not provide usable results when using primary results or a non-Recurrent Neural Network.

Based on the studies performed so far, we have concluded that a complex and diversified neural network must be constructed for the accurate future definition of Signal. In other words, a neural network of neural networks that examines the data set in several steps.

Depending on the above, we only tried to examine the change in the direction of the price movement. That is, for the oil price, we took the difference between the price of the actual day and the day before, to establish an increase or decrease in the price.

2.2. Analysis of Wall Street Journal articles

In the case of the analyzed journal, it was necessary to find one that met the following conditions:

- a leading magazine with a large readership, so it can have a real impact on the functioning of markets and speculative decision-making,

- has an archive file that is accessible and well structured,
- can be scraped, although this is more of a programming issue.

During the analysis of the websites, the Wall Street Journal (hereinafter: WSJ) proved to be the most suitable for the data collection to be carried out, so the scraper was run here.
In the course of the research, articles published by the WSJ between 2000 and 2020 were downloaded.
If a request is received from an IP address within a certain time, the website may classify the user as a bot and block it. Due to delays and pauses, the code line is slow enough to allow online activity to be considered a human activity.
Also, during the scraping, I connected to the Internet via VPN, so in case of downloading an IP address, the generation of another IP address that has not yet been banned can be set manually.
The entire sequence of the code line took 1604 hours, which means that it ran continuously for nearly 67 days. Calculated approximately, it reviewed and saved articles for an average of 401 minutes a month. The period 2000-2020, ie 21 years, represents a total of 330,435 articles. The resulting .xlsx file is 558 MB in size.
In the WSJ analysis, we analyzed the words of the articles alone, not in context.
We've highlighted keywords to tag our articles. That is, we estimated the extent to which the article is about the development of the price of oil or any event that may affect the price of oil. These are marked separately in the attached list.
In addition, we highlighted indicator words, which mark the quantitative occurrence in the articles submitted for analysis by the keywords. These give a hint about the future direction of the price, investor attitude, or market sentiments.

```
key_word_list = [

    'crude','opec','oil price','wti','crude oil',

    'increase', 'rise', 'rising', 'grow', 'optimism', 'enhance', 'expensive', 'climb', 'optimal', 'agreement', 'cooperation', 'solution', 'deal', 'bull', 'gain', 'demand', 'positive', 'decrease', 'bear', 'fall', 'low', 'cut', 'dramatically', 'pessimism', 'emergency', 'emerge', 'recession', 'collapse', 'negative', 'reduce',  'disagree', 'decline', 'cheap']
```

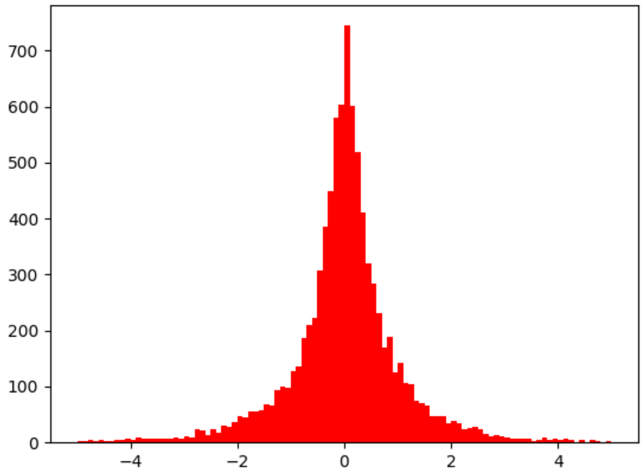
Current keywords are based on frequency, but we did not examine them for NLP for the present analysis, so the words are predefined by us. That is, this aspect of the results should be treated with caution, as we could have obtained other results by excluding or taking in some words. Respectively, word connections could have yielded even more sophisticated results. In any case, the current shortcomings can be addressed in part by a neural network. Also, depending on the subsequent results, we can evaluate the quality and relevance of the selected words.

3. RESULTS

The Artificial Neural Network was run with several settings.
Before describing the configurations, it is important to clarify that in the standardized data table for both oil prices and articles, the former 70% was used as the training set and the subsequent 30% as the test set. That is, at 70%, the neural network adjusted the neurons, reducing the margin of error by 500 epochs using the backpropagation method. The resulting neural network was run unchanged on the remaining 30%, was compared with the actual results, and with the extent to which the values predicted by the Neural Network corresponded to actual data.
We have changed the following or tested it for several values with one setting method:

- without deletion, ie every day is taken into account / deletion, ie days without an article with a minimum of 1 minimum indicator are deleted, so that no further analysis is performed,
- minimum accepted indicator words: 1, 2, 3, 4, 7, 10, 13, 16, 20, if the article does not have at least this many indicators, we will not analyze it further, thus increasing the analysis of the relevant articles;
- for the same day (t) and the previous day (t-1), ie practically the same day and the following day, thus preparing the one-day forecast;
- the magnitude of the difference between the results of the test set, ie the predicted change and the actual change: 0.02, 0.04, 0.07, 0.1, 0.15, 0.25, 0.4, 0.8, 1.0;
- stipulation of minimum change, i.e. the test set was considered only in the case, and the teaching set also examined only those where the real change and the change predicted by the neural network reached a certain level. We examined without a minimum expectation as well as a minimum change of 0.3 and 0.6;
- only a sign test, according to which the forecast predicts positive or negative values and how this relates to the actual exchange rate.

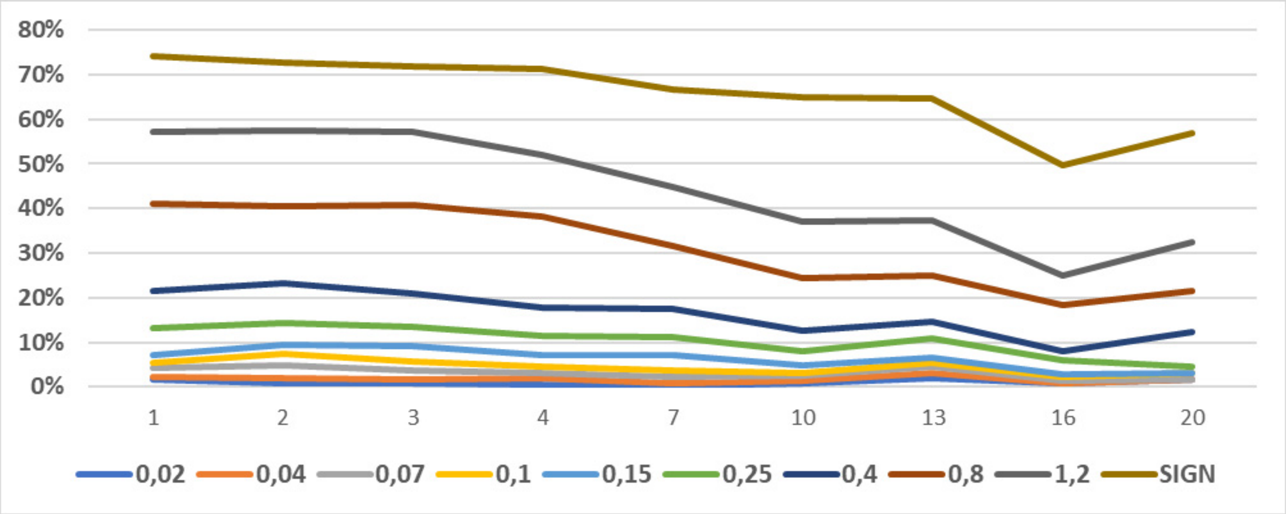
Before evaluating the results, it is important to know the extent of the changes, how volatile they are, and within what limits. This is shown in Picture 1.



Picture 1: The rate of change in WTI oil prices between 2000 and 2021
Source: own editing

It can be seen that the absolute value of the largest price change is 2, but rather rarely exceeds 1.6. In other words, we can say that the price of oil is volatile, but most of all it fluctuates around zero, it is not characterized by excessive fluctuations. This was to be expected given the stable and dominant position of the world market and the fact that price fluctuations have a major impact on many players and products in the economy. It can also be seen from the above figure that the positive and negative changes are almost the same, ie it is extremely important that the exchange rate will be heading at a given time.

Raising the indicators showed a consistent picture over the various runs. As we raised the indicator words, efficiency deteriorated. That is, increasing the number of indicator words cannot improve efficiency. This is presumably due to the extreme decline in the articles analyzed, which is squeezing the available Big Data source to such an extent that the Neural Network is no longer suitable for mapping real relationships and correlations.



Picture 1: Aggregate results of neural network runs by increasing indicator words as a function of efficiency by analyzing the different accepted maximum error limits separately
Source: own editing

Regarding the results of the analysis, the difference between the results of the same day and those of one day showed a slightly higher efficiency, but the result is not significant. Thus, both can be suitable for further analysis and prediction.

As expected, increasing the margin of error has greatly increased efficiency. The result can be considered accurate if it moves within 0.1 range, but it is around 5% or below it, which means that the present analysis does not give accurate values. The evaluation should take into account that too large a range already allows for a change of sign, hence it is possible that there will be a positive and negative change of result, which is a fatal mistake regarding the original purpose of the research.

Table 1: Neural Network analysis results

		indicators								
		1	2	3	4	7	10	13	16	20
max tolerance accepted	0,02	0,54%	1,06%	1,04%	1,12%	1,18%	0,55%	1,42%	0,00%	0,00%
	0,04	1,62%	1,91%	1,74%	1,95%	1,76%	1,65%	2,84%	0,00%	0,00%
	0,07	3,68%	4,03%	3,82%	3,63%	1,96%	2,20%	3,79%	0,00%	0,00%
	0,1	5,30%	5,73%	5,79%	5,44%	2,55%	2,20%	5,21%	0,83%	1,85%
	0,15	9,07%	8,70%	8,80%	7,53%	3,73%	3,85%	6,64%	1,67%	1,85%
	0,25	13,57%	14,12%	16,09%	10,74%	8,63%	7,42%	9,00%	5,83%	5,56%
	0,4	21,11%	21,97%	23,96%	16,32%	15,29%	13,19%	12,80%	9,17%	14,81%
	0,8	40,70%	40,23%	40,74%	35,98%	29,80%	26,92%	21,33%	19,17%	27,78%
	1,2	56,15%	57,64%	55,32%	51,60%	40,39%	37,36%	28,44%	26,67%	37,04%
	SIGN	74,03%	73,89%	72,57%	66,81%	65,69%	64,01%	54,03%	54,17%	59,26%

Source: own editing

In the case of runs with different settings, the most efficient one that can be used for investment decision making from the present research is the Sign analysis, i.e. the positive or negative deviation of the change from 0.

Overall, it produces results between 70-80%, which may be appropriate for decision making. Indicating entry and exit points.

Table 2: Neural Network analysis results

ERASE	MIN DIFF	ACCURACY
NO	0	79,77%
YES	0	78,43%
NO	0,3	74,08%
YES	0,3	74,03%
NO	0,6	68,15%
YES	0,6	66,30%

Source: own editing

4. DISCUSSION

This publication is the partial result of a larger study. The relationship and correlation can be demonstrated between the price and the content of the examined news items. That is, we can state (if not with absolute certainty) that there is a strong correlation between market price fluctuations and media opinion. It is a question of how accurately our forecast can be used after further thought, i.e. as a decision-making model.

The exact results are not accurate, this can be seen from the research. If you allow too high a margin of error, you will start to accept acceptable efficiency values, but this cannot be used for the reasons described.

In essence, the 80% value shown at the results will be a good basis for further research as well as investment decisions that can be used in part.

There are several ways to increase additional efficiency, including:

- application of Recurrent Neural Network, ie analysis of price volatility, including forecasting, confirmation with keyword research;
- Natural Language Processing, ie the examination of words and word phrases is not defined by the Neural Network, thus eliminating possible misuse of words

In the course of the research, the connection was established, and further efficiency gains are essential. Also, in anticipation of further research, it is important to state that the price is also influenced by macroeconomic factors, so it cannot be based solely and exclusively on press articles.

However, the hypothesis to continue with is that without making uncertain results, a decision-making model based on only openly available market information (in this case WSJ articles) can be created that can provide information on entry and exit points in prices with almost 100% results.

REFERENCES

Arshad, S., Rizvi, S. A., Haroon, O., Mehmood, F., & Gong, Q. (2021). Are oil prices efficient? *Economic Modelling, Vol. 96.*, 362-370. doi:10.1016/j.econmod.2020.03.018

EIA - Cushing, OK WTI Spot Price FOB (Dollars per Barrel). (2022.. január 5.). Forrás: https://www.eia.gov/dnav/pet/hist/RWTCD.htm

Aiyer, S., Niranjan, M., & Fallside, F. (1990). A theoretical investigation into the performance of the Hopfield model. *IEEE Transactions on Neural Networks, Volume: 1, Issue: 2,* 204-215. doi:10.1109/72.80232

Hopfield, J. J. (1982). Neural networks and physical systems with emergent. *Proc. Nat. Acad. Sci. Vol. 79 Issue 8*, 2554-2558. doi:10.1073/pnas.79.8.2554

Jain, A. K., Mohiuddin, K. M., & Mao, J. (1996). Artificial neural networks: a. *Computer Vol. 29. Issue 3.,* 31-44.

Rojas, R. (1996*). Neural Networks - A Systematic Introduction.* Berlin, New-York: Springer-Verlag.

Rabunal, J. R. (2005). *Artificial Neural Networks in Real-Life Applications*. IGI Global.

Soumya, C. V., & Ahmed, M. (2017). Artificial neural network based identification and classification of images of Bharatanatya gestures. *Innovative Mechanisms for Industry Applications (ICIMIA),* 162-166.

Keaton, L. L. (2017). *Consider the Community: Developing Predictive Linkages between Community Structure and Performance in Microbial Fuel Cells* Doctoral dissertation.

Farhat, N. H., Psaltis, D., Prata, A., & Paek, E. (1985). Optical implementation of the Hopfield model. *Appl Opt Volume 24.,* 1469-1476. doi:10.1364/AO.24.001469

Raza, K. (2017). *Prediction of Stock Market performance by using machine learning techniques. 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), 1-1.* doi:10.1109/ICIEECT.2017.7916583

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Comput Oper Res. Vol. 32. Issue 10.*, 2513-2522.