UDC 004.05

OSINT Time Series Forecasting Methods Analysis

A. Feher[1], D. Lande[1]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" , Kyiv, 03056, Ukraine*

Abstract

Time series forecasting is an important niche in the modern decision-making and tactics selection process, and in the context of OSINT technology, this approach can help predict events and allow for an effective response to them. For this purpose, LSTM, ARIMA, LPPL (JLS), N-gram were selected as time series forecasting methods, and their simple forms were implemented based on the time series of quantitative mentions of nato, himars, starlink and cyber threats statings obtained and generated using OSINT technology. Based on this, their overall effectiveness and the possibility of using them in combination with OSINT technology to form a forecast of the future were investigated.

Keywords: time-series, prediction, osint

Introduction

Business, finance, logistics, medicine, biology, and chemistry, use forecasting as one of the most applied methods of science that help to effectively solve typical problems and contribute to overall developments. At the same time, in the modern world, the latest neural network developments find their fits in various cybersecurity fields, such as threat intelligence, malware detection, and endpoint protection, which use probabilistic forecasting concepts for training, as well as show overall needs in the chosen topic.

Time-series forecasting methods as a scientific attitude use historical and current data to predict future values over a period of time or at a certain point in the future. By analysing the available data stored in the past, forecasting helps to understand future trends and allows you to respond to them in the most effective way.

In today's world, a well-designed forecasting system frees up hands and gives freedom in the field of the targeted application, even within the framework of national and cyber security. From the point of view of military and civilian security, such a system allows for the correct construction and adjustment of tactics and strategy at different time intervals in accordance with the forecasted events.

The task of the study is, first of all, to create a basis of the most effective forecasting methods for effective further research, and make qualitative comparisons between methods of its different nature. The methods themselves are analysed and used in conjunction with Open Source Intelligence (OSINT) technology to prove the application probability concept. The time series considered for the forecasting study represent quantitative collected information obtained using OSINT technologies.

Methodology

The selected time series for the study represents a complex dependence of the number of selected events obtained using OSINT technology on the time interval of one year. The selected event for analysis was presented as a dependence on the quantitative mentions characteristics of nato, himars, starlink, and cyber attacks statings in news, blogs, and articles all over the Internet on the corresponding time period using Infostream as a main source of statistics.

To create a comparative base, datasets were prepared in the manner only 334 days out of 365 were used to train the selected models, where the last month in a count of 31 days of the selected year was used as the predicted outcome values for further analysis.

For forecasting time series need to classify that time series under study are predictable. Proposed to use the Hurst exponent to evaluate the applied predictability of chosen datasets, as the prime technique which provides a measure for long-term memory and fractality to evaluate the stationarity of time series, where the value of stationarity $\boldsymbol{H}$ ranges between 0 and 1. Measure characteristics depend on mean-reverting strength, if the strength of the trend approaches to value of $\boldsymbol{H} > 0.5$ is described as persistent series, and vice versa if the trend approaches $\boldsymbol{H} < 0.5$ it describes as non-persistent, and not applicable to forecasting techniques.

Estimating method for the Hurst exponent was calculated with the expected Rescaled Range Analysis (R/S analysis) value [1] which is associated with the normalized spread coefficient. This method is based on the cumulative sum of the time series, denoted by $\boldsymbol{R}$, and the standard deviation sum, denoted by $\boldsymbol{S}$, the length of the time series, denoted by $\boldsymbol{N}$, and can be calculated with a linear regression equation (1).

$$\boldsymbol{R/S} = (\boldsymbol{N}/2)^{\boldsymbol{H}} \tag{1}$$

In turn main modern approaches to forecasting are considered to be the following: neural network, statistical, econometric, and linguistic. Each of them is actively used in their respective industries, and in some cases, a combination of several approaches or tuning modifications is used to obtain the most relevant values of needs.

As typical modern representatives of the described approaches, the following methods have been chosen to study time series and build forecasts series corresponding to them:

- Long Short-Term Memory (LSTM) as the most common neural network method;
- Autoregressive Integrated Moving Average (ARIMA) as the most widely used statistical method;
- Log Periodic Power Law (LPPL) or Johansen-Ledoit-Sornette (JLS) as an econometric method, which is subject to criticism and is not popular, but is used in some cases;
- N-gram as a linguistic one, which is already quite strongly implemented in modern life.

Depending on the selected datasets, chosen forecasting models can predict the quantitative characteristics of selected events based on corresponding values and there is a need for a possibility to calculate the relation level between the real and predicted data.

Conducted accuracy measurement used an under probationary method for the forecasting techniques, where expression of its calculation is relationships between sets, for this purpose observed and took into consideration multiple practices to manage correlation, where the most suitable were Spearman correlation for measuring monotonic relationships and Pearson correlation coefficient (PCC) for measuring linear relationships to achieve conclusive analysis of predicted values preciseness. According to the similarity between gathered measurement values of both mentioned methods was suggested that the degree of linear relationships describes interrelation more transparently, and PCC was selected.

Correlation calculation is based on covariance which also could serve as an indicator of variety metric between variables in a linear relationship nevertheless it can go infinitely positive or negative is not so feasible in practice, that's why chose the PCC which additionally stands on standard deviations estimating and occur as a more precise and vivid measure for accuracy. The correlation coefficient is an indicator of the degree of linear dependence between variables in a range of [-1,1] by normalizing covariance with mentioned standard deviations, where +1 is a perfect direct linear relationship, and -1 is a perfect inverse linear relationship [2]. Correlation is calculated based on the equation (2), where xi, yi - actual values and predicted values, and xx, yy, - sample mean of actual and predicted values.

$$r = \frac{\Sigma\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\Sigma\left(x_i - \overline{x}\right)^2 \Sigma\left(y_i - \overline{y}\right)^2}} \tag{2}$$

LSTM

LSTM is a kind of recurrent neural network model, with the difference that LSTM can handle long time series of data. In addition, the conventional recurrent model has a vanishing gradient problem for long data sequences, while LSTM can prevent this problem during training and perform qualitative results, in terms of three gates architectures. These three gates are analog gates and their neural cells are based on the sigmoid function [3] which works in the range 0 to 1. Where an input gate determines whether or not to let the new input in, a forget gate deletes previously stored information, and the output gate decides what information is considered to output.

The model can recall previous long-time series of data [4] and has automatic controls to keep relevant features or discard irrelevant features. It is because of these factors that LSTM was chosen among other recurrent methods as a method for the study.

For LSTM, we used its single-layer configuration with 32 units, and the Adam optimizer as an extension to stochastic gradient descent which gave more relevant results, with batch size and epoch values of 512 which defines a number of predictions at the time, the output data was (inverse) transformed by normalization to obtain the predicted series.

ARIMA

ARIMA is a time series analysis model which stands on lags and shifts in the historical data to uncover patterns such as sequences, moving averages, and seasonality to predict future data. It is an autoregressive integrated moving average model, where the AR part shows that the time series is regressed on its own past data and describes a stochastic process for the weighted sum of its previous values and a white noise error. The MA part shows that the forecast error is a linear combination of past corresponding errors and describes a stochastic process for the weighted sum of a white noise error from current and previous periods. The I part shows that the data values have been replaced by different values of order d to obtain stationary data, which is a requirement of the ARIMA approach.

It is because of this complexity that the ARIMA model is effective in re-examining past data using this combined learning approach and helps to effectively predict future points in the time series [5]. This attitude creates a base of popularity for the method and its practical value.

For ARIMA, we used its one-layer configuration with $p = 33$ which defines the number of lag observations included in the model, $d = 2$ which defines the number of times that the raw observations are differenced as a degree of differencing, $q = 0$ which defines the size of the moving average window, the values of which were determined empirically according to the more relevant output forecast values.

LPPL

The LPPL or Johansen-Ledoyt-Sornett (JLS) model – attempts to diagnose, time, and predict the end of financial bubbles, a common term in the financial industry for crisis points when the majority of participants lose confidence during speculative growth. The primary part of the JLS model referred to Martingale's discrete-time stochastic process with its probability theory, on which it is stationed.

Despite the widespread criticism [6], the creators of the model provide a motivation based on some natural assumptions, including risk-neutral assets, rational expectations, local self-reinforcing imitation, and probabilistic critical moments for the algorithm to calculate the stages of bubble development directly [7] with a simple equation which mostly depends on Dragon King and Black Swan event concepts and behavioral finance theory. This way, can see how the chosen forecasting algorithm works with an atypical for it time series.

With mentioned criticism, there are founded various approaches for studying LPPL (JSL), original 2-step nonlinear optimization, 3-step estimation optimization, backpropagation implementation, and genetic algorithms. The study used its modification using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), which gives a more varied and relevant forecasted series with keeping the original idea.

N-gram

N-grams represent a continuous sequence of N elements as a Markov process from a given set of texts. The N-grams technique has found its main application in the field of probabilistic language models. They estimate the probability of the next element in a sequence of words, and this is the basis of the theoretical approach to assume study time series forecasting. After all n-gram models are also often criticized in a case of lack of any explicit representation of long-range dependencies and their inefficiency and narrowness in other applied research subjects.

This approach to language modeling estimates a close relationship between the position of each element in the string, calculating the occurrence of the next word in relation to the previous one and the frequency of their occurrence. In a broad sense, these elements do not necessarily mean strings of words, they can also be phonemes, syllables, or letters [8], depending on what exactly is required, and it is thanks to this flexibility that the work was able to be based on numeric time series as well.

There is an additional variation in modeling by creating semantically connected elements in turn, in this paper, the unigram was studied as N-gram with one connection inside, to provide a complete forecast of 31 days, with other values of N the model could not produce a chain of values with a length of 31 values, and a simple general type of tokenization of all elements was used.

Results and Discussion

The software was developed for each method, and the time series was adjusted to obtain the predicted results. The graphs shown in Figure 1 of actual (real) and predicted values were modeled according to the dataset of chosen time series, and the processes were repeated to obtain the most relevant predicted series.

Neural network and statistical approaches proved to be the most effective for forecasts, while econometric and linguistic methods proved to be rather limited in their use in forecasting such time series.

LSTM

The LSTM method is quite flexible and can be easily adjusted to the specifics of the time series, due to its complexity, the method works stably, without fail, and the predicted results are quite close to the real ones (Fig. 1). It is also possible to adjust additional hyper-parameters [9], so it is possible to create a multilayer model with stronger rejection, which can give more accurate predicted results as well as examine different configuration values with an increased number of hidden layers.

ARIMA

ARIMA was a good choice, it is less flexible in use, but with the correct selection of parameters p, d, q it makes its forecasts quite accurately according to different kinds of time series [10], among the selected options it showed itself to be the best (Fig. 1). In spite of this approach also there is way to increase accuracy including seasonality and exogenous variables like Seasonal ARIMA (SARIMA), ARIMA with eXogenous factors (ARIMAX), combination of it - SARIMAX or using neural network hybrid models such as SARIMA Back Propagation (SARIMABP), or take inspiration from other seasoning and autoregression models for future considerations.

LPPL

In another way, LPPL performs rather poorly as a method for forecasting time series, which is not surprising due to its narrow focus on solving other mentioned problems (Fig. 1). The model is still evolving over time, partly in response to valid criticism, and in the course of the study was found that the strategy of evolutionary adaptation of the covariance matrix CMA-ES is a good improvement that allows for more accurate results, but despite this, when using generative algorithms such as CMA-ES to improve the forecast, the complexity of the calculation itself increases proportionally. It turned out that the calculation of individual large numerical values is also problematic, which requires taking their logarithmic representation, which can also affect the distortion of the forecast.

N-gram

The N-gram model presented a rather limited version of time series forecasting due to the limited number of previous possible values according to which the forecast can take them (Fig. 1). That is, considering this method within the framework of non-stationary series, the forecast is limited by the threshold values of the time series and cannot go beyond it, which reduces its accuracy. Therefore, in studying time series, there is wide room for improvement when using the model with the N-1 algorithm, in which the forecast distortion at short intervals will be much smaller and gradually graduated with respect to time, and the addition of a recurrent component that can increase the accuracy at longer time intervals. It is also worth noting the creation of joint or separate dictionaries for different series, which will increase the accuracy of joint series if there is an appropriate semantic correlation, but vice versa in the absence of such correlations.

Calculated results for Hurst exponent values are illustrated in Table 1. where a higher number reflects a higher forecastability, as well as the PCC values in Table 2 which to seek sufficiency, were applied multiple iterations of proposed forecasting methods and where also a higher value describes a perfectness linear relationship between real and predicted data.

Table 1. Hurst exponent

Starlink	Cyber Attacks	Himars	Nato
0.81127	0.86121	0.90728	0.95844

Table 2. PCC

	Starlink	Cyber Attacks	Himars	Nato
LSTM	0.243	0.272	0.358	0.357
ARIMA	0.286	0.373	0.388	0.427
LPPL	0.185	0.176	0.212	0.286
N-gram	0.209	0.214	0.266	0.305

Laying on gathered results despite their appearance pretended to be unpleasant all methods display positive correlation, state that due to a small-scale sample study and its sensitivity to outliers, it can’t be definitely marked and perceived as an inferior measuring approach to be applied in forecast kind of researches, and a statement for its suitability for such kind of studies is open as well as debatable, and admitting trends it shows can be reasonably assumed that forecast accuracy significantly depends on the Hurst exponent values of chosen training datasets.

Due to the diversity of datasets, all across rational space, may be concluded that the level of stationarity enacts a major part in forecasting accuracy along with properly configured prediction models.

Conclusion

Based on the empirical part, it can be noted that each of the considered methods satisfies the task applied to persistent time series, despite the low accuracy of effective time series forecasting of such models as LPPL and N-gram, they provide much more creative space for further study and optimisation. In turn, LSTM and ARIMA models have proved to be quite effective, so it is not surprising that these models and their approaches are dominant in terms of time series forecasting.

Due to the efforts put in place, a basis for further study has been placed, proven forecasting various types of events obtained from open sources and in particular the models themselves. In the context of this study, having the means of automated OSINT data collection, it is possible to confirm the effectiveness of their use for building quantitative predictive scenarios for the future we live in.

References

1. *Lande D.* Fractal properties of thematic information flows from the Internet // ISSN 1560-9189 Registration, storage and processing of data, 2006, Vol. 8, No. 2. — 2006. — P. 2–4.
2. *Friday Zinzendoff Okwonu Bolaji Laro Asaju F. I. A.* Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods. — 2020. — DOI: 10.1088/1757-899X/917/1/012065.
3. *Anita Yadav C K Jha A. S.* Optimizing LSTM for time series prediction in Indian stock market. — 2020. — DOI: 0.1016/j.procs.2020.03.257.
4. *Sudriani Y., Ridwansyah I., Rustini H. A.* Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia. — 2019. — DOI: 10.1088/1755-1315/299/1/012037.
5. *Brownlee J.* Introduction to Time Series Forecasting with Python. — 1st ed. — 2020. — 365 p.
6. *Fantazzini D., Geraskin P.* Everything You Always Wanted to Know about Log Periodic Power Laws for Bubble Modelling but Were Afraid to Ask // European Journal of Finance. — 2011. — Vol. 19. — P. 11–13. — DOI: 10.1080/1351847X.2011.601657.
7. *Shu M., Zhu W.* Diagnosis and Prediction of the 2015 Chinese Stock Market Bubble. — 2019. — arXiv: 1905.09633 [q-fin.ST].
8. *Jurafsky D., Martin J. H.* Speech and Language Processing. — 3rd ed. — 2023. — 636 p.
9. *Staudemeyer R. C., Morris E. R.* Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. — 2019. — arXiv: 1909.09586 [cs.NE].
10. *Charisios Christodoulos Christos Michalakelis D. V.* Forecasting with limited data: Combining ARIMA and diffusion models. — 2010. — DOI: 10.1016/j.techfore.2010.01.009.

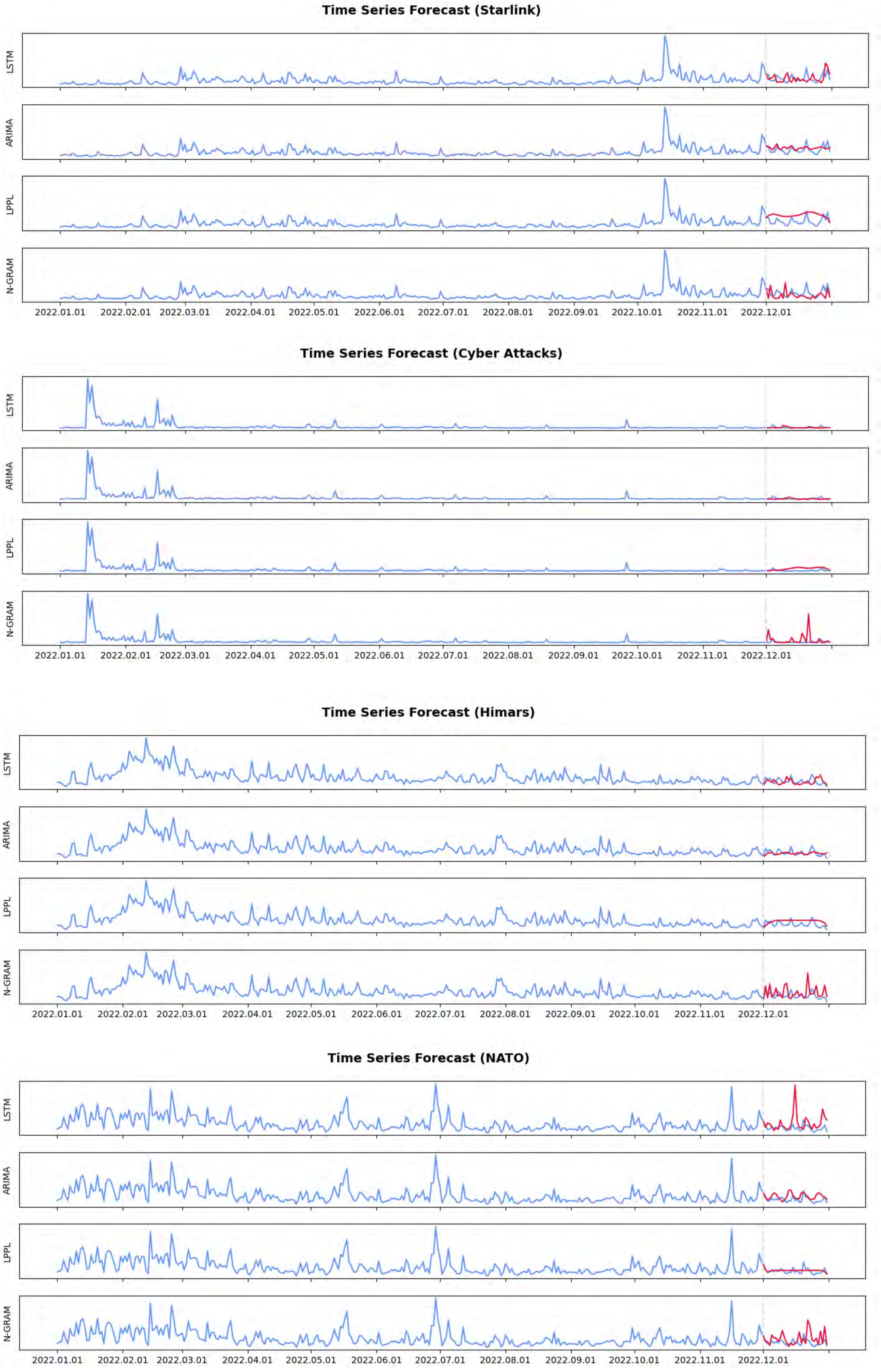

Figure 1. The rows show applied forecast methods by chosen dataset, the segment for the last month – grey cut, the actual time series – blue, the predicted time series – red