Impact of social media news on stock market index

Jakub Ajchel
Data Science Lab

Abstract

This work focuses on forecasting prices in the stock market with the use of traditional time series models and machine learning algorithms. Author's main aim is to check if stock prices are predictable, which would mean that the traders can obtain profits and thus financial markets are not efficient. It is tested if machine learning algorithms, especially LSTM, give better results in time series forecasting than the traditional ARIMA model. In addition, the author verifies the impact of additional information, i.e. the sentiment of the social media news, on the accuracy of stock prices predictions. Three research hypotheses that will be empirically verified are presented on the right-hand side:

Contact with the author: e-mail: jakubajchel@gmail.com; www.linkedin.com/in/jakubajchel; https://dslab.wne.uw.edu.pl

Hypotheses

1. **Financial markets are not efficient. Historical data can be used to predict the current price of an asset.**
2. **Machine learning algorithms like neural networks outperform traditional linear time series models like ARIMA in stock prices forecasting.**
3. **Additional information from social media news (sentiment analysis) increase the accuracy of stock market predictions.**

Theoretical introduction

One of the most important concepts of time series analysis is stationarity. Time series is stationary if its statistical properties do not change over time. Stationarity might be tested using Dickey-Fuller (DF) or Augmented Dickey-Fuller (ADF) tests. Stationary processes are well understood however non-stationary series are more common. If the expected value of the non-stationary stochastic process changes over time, we can call such a process a time series with a deterministic trend. Applying regression analysis on time series that contain trends might lead to spuriously good results. Another concept that is associated with stationarity is integration. We say that the time series is integrated of order d, if it is not stationary, but we can transform it into stationary by differencing it d times. By differencing a time series, we can remove the trend.
Traditionally, a theory of finance presented financial market as efficient, i.e. prices of financial assets are unpredictable and follow a random walk (succession of random steps). Statistically, it means that each successive price is independent of the previous ones. Eugene Fama formulated the Efficient Market Hypothesis (EMH), that states that the current prices on the market are the correct ones. Any past information is already reflected in the price. Fama divided empirical tests of efficiency into 3 different forms that refer to the information set used in the statement "prices reflect all available information" (emh2): weak-form, semi-strong and strong-form efficiency.
Even after more than 50 years of research, the consensus has not been reached in terms of the presence or absence of the validity of EMH. However, multiple researchers claims, that trying to predict stock market prices based on historic prices only is not sufficient. External stimuli should be also included, what is done in this work.
To do that, author of this work focused on sentiment analysis of tweets – their impact on stock market predictability was tested. For better results of sentiment analysis, the author has used GloVe, pre-trained log-bilinear regression model for the unsupervised learning of word representations. It outperforms other models on word analogy, word similarity or named entity recognition tasks. No research regarding the stock market prediction that uses sentiment analysis with GloVe has been found by the author.
What is more, the added value of this work is the methodology of combining Twitter sentiment and stock prices history. In most works, all tweets from the given day are taken into account. Author of this work focused only on tweets from users that are specialized on Oil, Gas and Energy markets. Then, their impact on the whole S&P500 index, but also impact on S&P Oil & Gas Exploration & Production index was checked, what is an innovative attitude to the research.

Methodology

The simplest class of time series model is that of the Moving Average (MA) process (output variable depends linearly on the current and various past values of a stochastic term) and Autoregressive (AR) process (current value of the variable depends only on the previous values of that variable plus an error term). We can combine them into ARMA model, that can be generalized into the ARIMA model (AutoRegressive Integrated Moving Average). What differentiates them, is that they can be used in cases where data is non-stationary. Initial differencing step can be applied one or more times to eliminate non-stationarity. Different ARIMA models can be characterized using 3 parameters – p (number of lagged observations included in the model; called the lag order), d (number of times that the raw observations are differenced; called the degree of differencing), q (the size of the moving average window; called the order of moving average).
ARIMAX (ARIMA model with an exogenous variable) is an ARIMA model with an additional independent variable. It is suitable for analysis where there are additional explanatory variables. Using ARIMAX enables author to add extra variables related to sentiment analysis of social media news.
Among all algorithms, Artificial Neural Networks (ANNs) is the most commonly used technique in machine learning stock prediction. Unfortunately, simple ANNs quite often suffer from the problem of overfitting. That is why the author has used Long Short-Term Memory (LSTM) model in this work. This is a kind of a recurrent neural network (RNN), but in opposite to standard RNN, LSTM has feedback connections. Instead of neurons, LSTM networks have memory blocks (units) that are connected into layers.
The most common architecture of LSTM (presented in Figure 1) is composed of a cell/block (the memory part) and three "regulators", called gates, of the flow of information inside the LSTM unit. The cell is responsible mainly for keeping track of the dependencies between the elements in the input sequence. Gates are responsible for managing the block's state and output. There are following gates in LSTM memory cell:

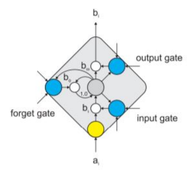

Fig. 1: Example of LSTM cell

- Forget Gate – decides what information to discard from the unit and how to use information stored in memory,
- Input Gate – regulates which values from the input to update the memory state by involving an activation function,
- Output Gate – decides what to output based on input and the memory of the unit. It selects useful information from the current cell state and showing it out as an output.

As there were multiple different emotions to be recognized, Softmax activation function was used. Softmax regression is a kind of logistic regression that normalizes an input value into a vector of values that follows a probability distribution that sums up to 1. Softmax turns the numeric output of the last hidden layer of a neural network into probabilities.
Different models should be compared with each other to assess which of them predicts the future most accurate. Author decided to use the Root-Mean-Square Error (RMSE). RMSE measures the differences or residuals between actual and predicted values. This metric compares prediction errors of models for particular data and not between datasets.

Empirical data used in research

In the first part of the research, data from the American Stock Market was used. The author extracted historical daily S&P 500 commodity price index (GSPC) from range 01.2014-12.2019 from the Yahoo Finance Website. "Close" price was chosen as the only feature of financial time series to be fed into models. The whole set was split (in a chronological way) into two subsets: training and test datasets with the 70/30 ratio. This data is plotted in figure 2. The plot is similar to a random walk with positive drift. The data is non-stationary.
To test the impact of the Twitter on stock prices, tweets from industry Twitter accounts like Oil And Gas Investor, Energy, Oil & Gas, Oil & Gas Journal or EY Oil & Gas etc. were used.
As it was impossible to find the dataset based on the Twitter accounts mentioned, author used Transfer Learning technique. It means that Machine Learning model for Sentiment Analysis was trained on tweets from Kaggle competition for identifying emotions in text dataset, and then this model was used to assess emotions of new tweets. The original dataset for sentiment analysis contains 40,000 tweets classified into 13 emotion classes. Some classes are very similar (e.g. fun and happiness), so they were combined and re-labelled with four emotions only: happiness, sadness, neutrality and anger. After training a model on that data, it was used to assess emotions on tweets from the industry accounts. They were gathered using Twitter API. 34733 tweets from a period 23.11.2009 – 16.08.2020 were collected. After predicting their tone, tweets from the same day were grouped and the average mood of that day was calculated. Finally, this average mood was combined with a dataset that contains stock prices.

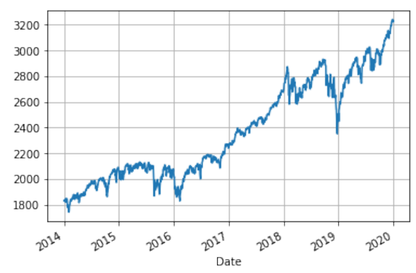

Fig. 2: Stock price of S&P 500 in years 2014-2019

Empirical analysis

The analysis consisted of 3 steps. In the beginning, different algorithms (ARIMA models, linear regression and neural networks) were compared as time series predictors. The second research hypothesis states that ML algorithms produce more accurate forecast than ARIMA models. In the next part, the sentiment analysis of tweets was conducted. In that part, also multiple algorithms were tested (logistic regression, LSTM and XGBoost). Finally, the best model from the second part was used to assess the sentiment of the tweets from the industry of oil, gas and energy. This sentiment was used as additional information in models from the first part of the analysis. The third research hypothesis states that this additional information will improve the performance of the models.

Time series modelling

In the first stage, one-step-ahead forecasts will be tested. Then, algorithms in the longer horizon will be tested (multi-step forecasts). In the longer horizon, values for the whole test set (30% so 453 observations) will be assessed. Such a long horizon should not be forecasted using ARIMA (number of observations to be forecasted should not exceed values of parameters p and q), but it was done purposely by the author to show the difference between ARIMA and other models.
Process of building ARIMA model followed the Box-Jenkins approach. Arima models with different parameters were tested and based on AIC criterion, ARIMA(4,1,4) was chosen as the best model. It was tested for multi-step forecasts as well. The model was not re-estimated so the first value that was estimated was the same as the estimation from one-step-ahead forecast.
Then, LSTM was tested on the same dataset. LSTM can hold and learn from a long sequence of observations - it is designed for longer predictions than one-step, however, to compare in fair conditions, one-step LSTM prediction was assessed. Dataset was scaled to [-1, 1], as this is required by LSTMs. Different networks were compared and the one with two layers of 50 neurons, one Dropout layer (with dropout parameter equals to 0.2) and Dense layer appeared to be the best. The "mean squared error" and "Adam" are used as the loss function and the optimization algorithm, respectively. LSTM was also tested for multi-step forecasts using closing prices from the whole range of dates from the test set.
The results of the experiments are presented in the table 1:

Model	RMSE for one-step prediction	RMSE for multi-step prediction
ARIMA	19.32	152.11
LSTM	24.94	119,30

Tab. 1: Results of models comparison for the same dataset, but for one-step and multi-step prediction

ARIMA yields better results in forecasting for the short term, while LSTM yields better results for long term modelling. Taking into account all the limits of ARIMA, it can be concluded that the LSTM model is promising in the field of stock prices forecasting and able to outperform the ARIMA model, what proves the second hypothesis of the work.

Sentiment Analysis

The first step of sentiment analysis was text cleaning. Then, tokenization was conducted. Words from each sentence are combined to create a vocabulary of all the unique words in the sentences. Then, all tweets will be turned into a sequence of integers. Each number will correspond to one word from the dictionary created. The last step of tweets preparation was to truncate and pad the input sequences so that they are all in the same length for modelling. This is considered as a Bag-of-words (BOW) model, which is a common way in NLP to create vectors out of the text. Each document is represented as a vector. In the last step, data was divided into training set (90%) and test set (10%).

Model	Logistic Regression	LSTM	Bidirectional LSTM	CNN+LSTM	Bidirect. LSTM with GloVe	XGBoost
Accuracy	34.83%	55.39%	55.55%	55.70%	55.98%	49.58%

Tab. 2: Summary of models performance for sentiment analysis

In the modelling part the following models were tested: logistic regression, LSTMs (with different optimizers, number of learning rates and, nodes and layers), bidirectional LSTMs (with different dropout, number of learning rates and nodes), combinations of CNNs with bidirectional LSTMs (with different kernel size, number of LSTM nodes and filters), bidirectional LSTM with GloVe and XGBoost. In the table 2, results of accuracy on validation set for different models can be found – bidirectional LSTM model with 32 nodes, GloVe embedding, SpatialDropout equals to 0.2 and Adam optimizer with learning rate equals to 0.001 was chosen as the best one. Model accuracy on the test set is: 53.90%

Time series modelling with additional data from Twitter

Finally, models for time series prediction will be combined with predictions given by the LSTM model for natural language modelling. The following models will be tested: ARIMAX, LSTM with 1 lag (simple neural network with one hidden layer with 50 neurons), LSTM with 4 lags (the same network as in the previous point).
It is important to mention, that in this stage, neural networks were not tuned. As very simple LSTM is used, its result cannot be directly compared with the result achieved by more complex LSTM from the previous section. What is more, data from a limited range of dates was used (only years 2016-2019) as in the previous period, a number of tweets were not sufficient. It means that results from this chapter can be compared between each other, but they cannot be compared with results from the section about Time-series modelling.
In the beginning, sentiment of tweets in each day was assessed by the best algorithm from the previous section. Then, it was combined with financial data. ARIMA(4,1,4) and ARIMAX(4,1,4) were first models to be tested. In the next step, LSTMs with 1 lag and 4 lags were built. In all models, mean absolute error as loss and Adam as optimizer were used. All results are summarized in the table 3:

Model	Dataset with stock prices only	Dataset with additional information from Twitter
ARIMA/ARIMAX	25.16	25.05
LSTM with 1 lag	31.83	28.25
LSTM with 4 lags	27,75	19.32

Tab. 3: Results of models comparison for datasets with and without additional information with average mood of tweets. S&P500 Index

From the Table 3 it can be seen that additional information from Twitter always improves the results of the model. These results prove the third hypothesis of the work that social media news improves the accuracy of stock prices predictions.
Then, impact of the Twitter news on Oil, Gas and Energy influenced S&P Oil & Gas Exploration & Production Index was checked. These results will be compared to the impact of this news on standard S&P500 Index. Only LSTM with 4 lags was checked, as the impact of additional information from social media has the biggest impact on this model. The RMSE between models results for different datasets cannot be compared directly. However, the percentage change in RMSE can be compared. LSTM with 4 lags for the dataset of S&P Oil & Gas Exploration & Production Index (in the same range of dates - September 2016 - December 2019) gave RMSE equals to 13.24. Then, the sentiment of Twitter news was added and RMSE decreased to 3.59. It means the change of 73%, that is much more than in case of standard S&P500 Index. The result is consistent with the expectations as news from the industry have a much bigger impact on the index chosen than on the general S&P500 Index. It also confirmed the hypothesis about the positive impact of social media news on stock market predictability.

Summary

To confirm the hypotheses, multiple techniques and algorithms were used. For stock prices predictions author has used ARIMA and ARIMAX, but also neural networks - LSTM networks. Multiple algorithms were also tested for sentiment analysis - Logistic Regression, multiple variations of LSTMs and XGBoost. The author has used advanced methods of boosting, hyperparameter tuning and transfer learning to achieve better and more reliable results. The usefulness of LSTM neural networks in different types of problems was confirmed. It appeared to be the best model for both time-series prediction (with additional information) and sentiment analysis of short messages. What is more, during the research, all three hypotheses were confirmed:

- Financial markets appeared not to be efficient. They can be predicted based on historical data.
- Long Short-Term Memory neural networks appeared to outperform ARIMA for multi-step predictions of time-series. On the other hand, ARIMA achieved better results for one-step predictions.
- Additional information about the sentiment of news from Twitter increased the accuracy of the predictions.