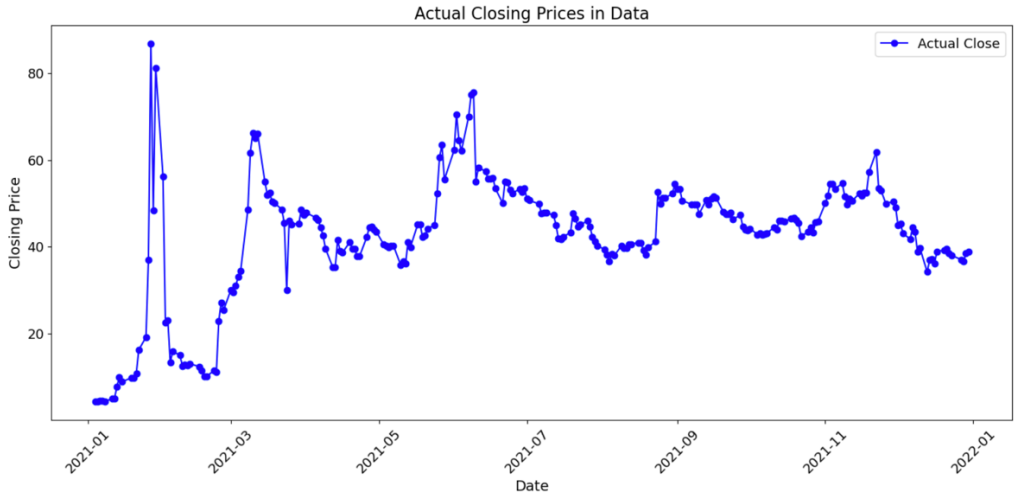
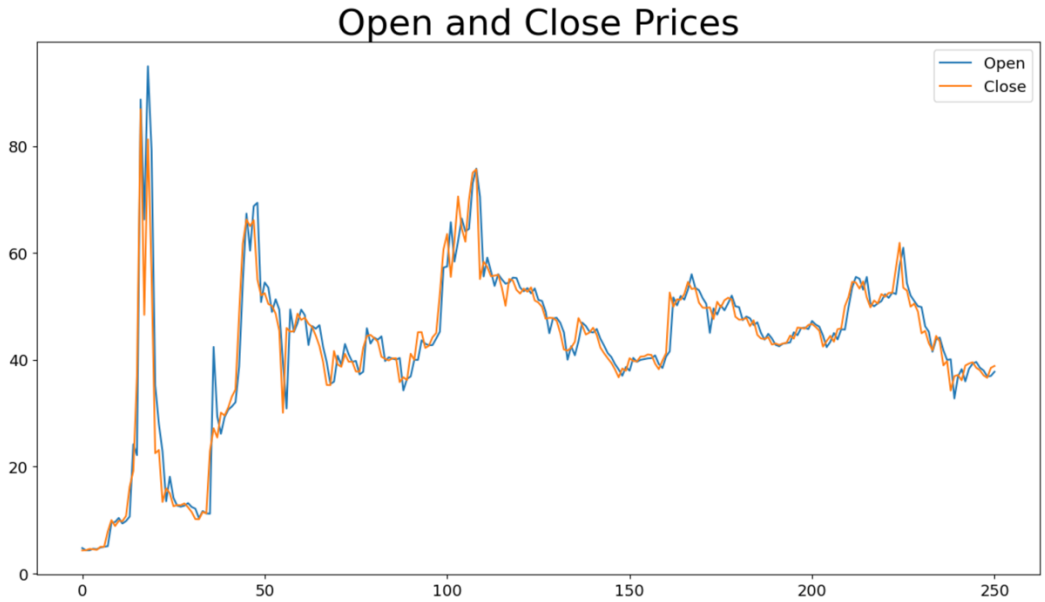
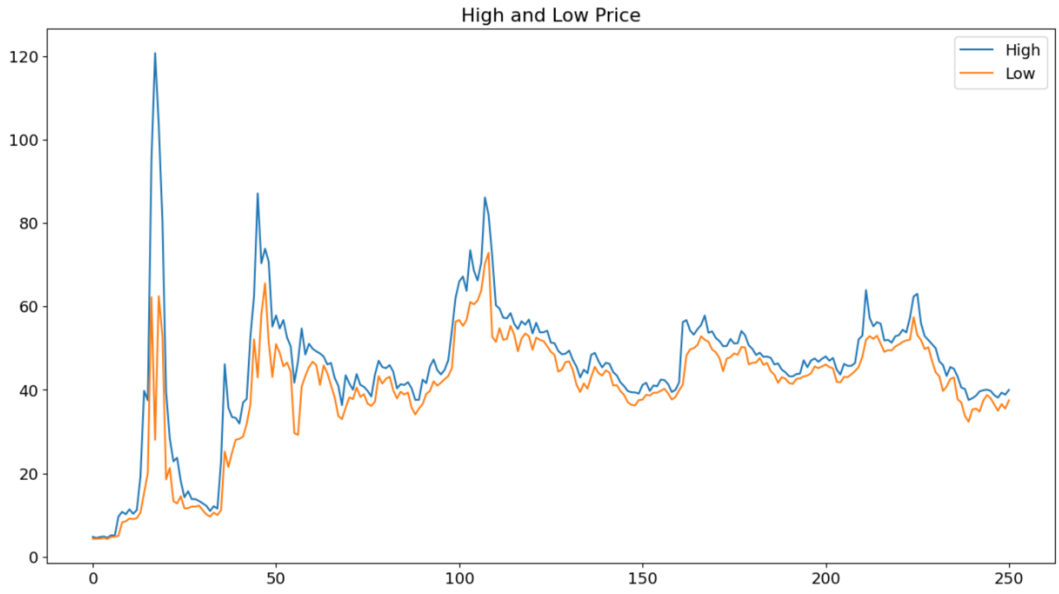
Muhammad Asghar
Applications of NLX and Large Language Models
Assignment 1 Report GME Short Squeeze

GameStop stock experimented a rise in popularity in 2021. This was due to social media activity where users attempted to “short squeeze” the stock. This project looks into the sentiment behind that phenomenon based off reddit comment history as well as GME stock price in 2021. The data comes from Yahoo Finance from 01-04-2021 to 12-31-2021 and from Harvard Dataverse Han, Jing, 2022, "Reddit Dataset on Meme Stock: GameStop".

To begin, I started with basic exploratory data analysis on general stock prices in 2021.



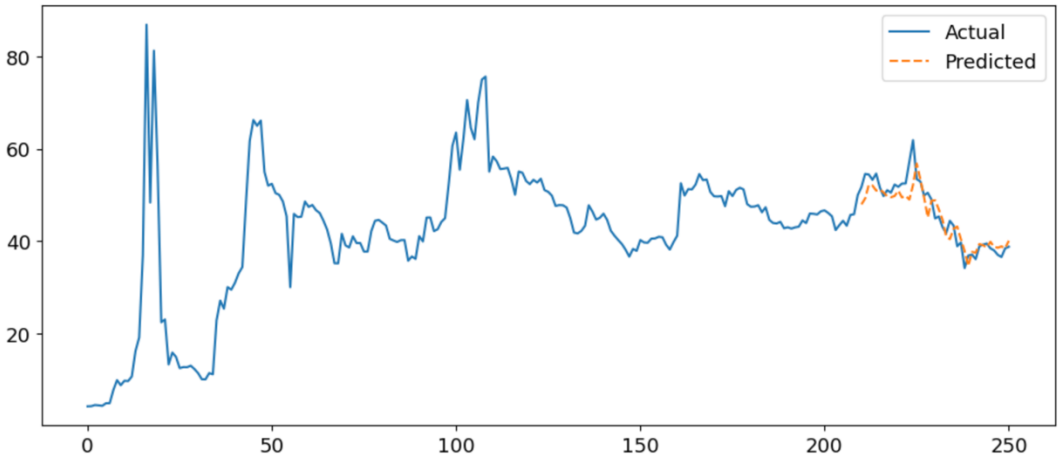


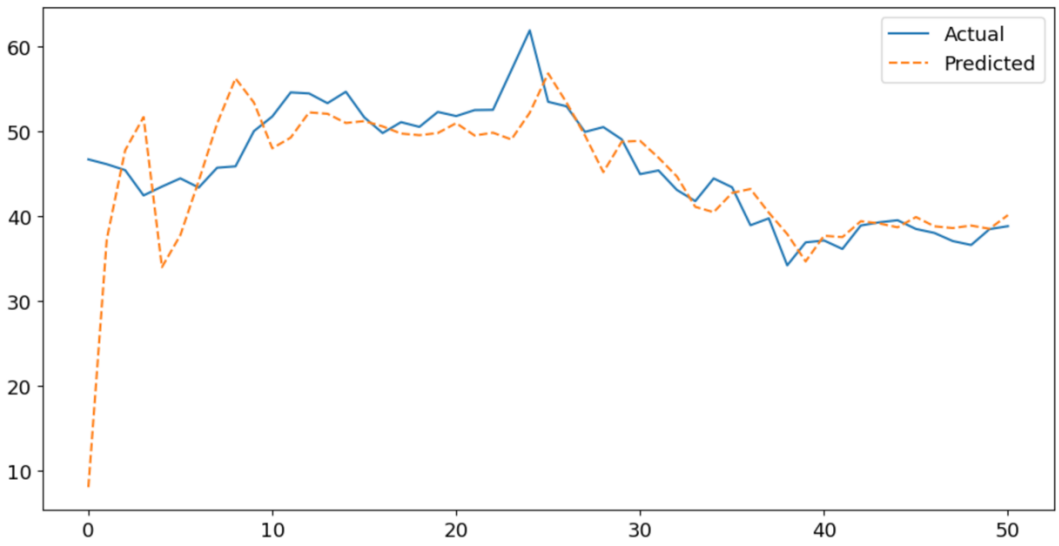


Generally, there were many spikes in the data and this must be investigated further. To analyze the data, I used a Recurrent Neural Network and Long Short-Term Memory approach to predict the stock prices after May 31, 2021. The training was done prior to May 31, and the test was after May 31.

With RNN I have the following architecture. A one-dimensional convolutional layer (Conv1D), which applies filters to the input data to extract relevant features. It uses 5 filters, each with a kernel size of 3 (i.e., a sliding window of 3 time steps). The activation function used is ReLU (Rectified Linear Unit). Then I use a Simple Recurrent Neural Network (SimpleRNN) layer to the model. This layer processes sequential data (such as time series) by maintaining an internal state. It has 32 hidden units (neurons). I then include a dropout layer to prevent overfitting during training, randomly setting 20% of the input units to 0 at each update. Then comes a fully connected (dense) layer with a single output unit. This layer performs a linear transformation on the input data. Lastly, I apply a linear activation function to the output of the previous layer. In this case, it maintains the output as-is (no activation function).

I use the mean squared error as the loss function and the adam optimizer. Then to fit the model, I use 100 epochs, and a batch size of 18. I use 10% for the validation split.





As for performance, I calculated the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

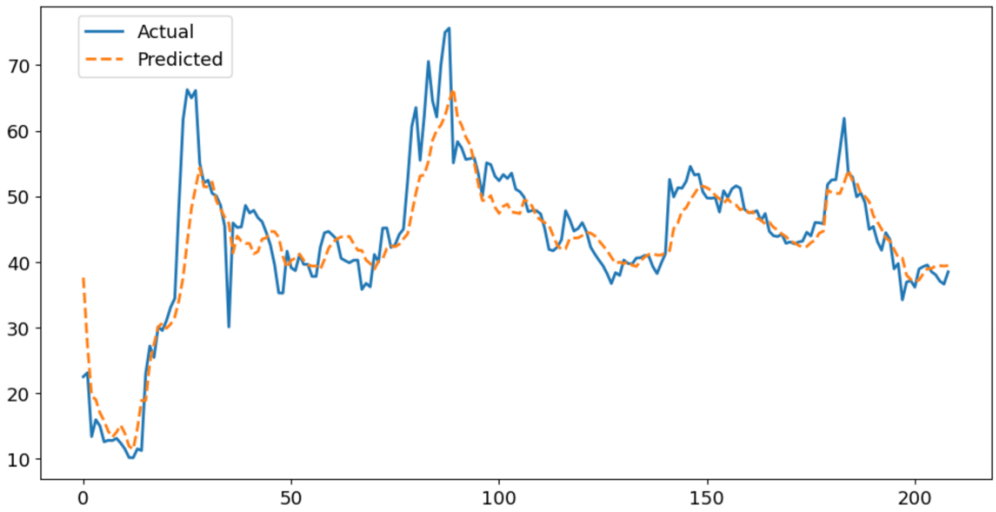
As for the performance metrics I get the following:

Train MAE: 3.9%
Train RMSE: 7.1%
Train MSE: 0.5%

Test MAE: 4.6%
Test RMSE: 9.0%
Test MSE: 0.5%

It performed very well but there is room for improvement.

For the LSTM model I set a window size of 20 (assuming 5 trading days). I set the number of units (neurons) to 50 and use the ReLU activation function. I then incorporate a dropout layer during training, randomly setting 20% of the input units to 0 at each update. Then I add a dense layer with the linear activation function. For the loss, I compute the mean squared error and have the optimizer set to adam.



As for the performance metrics I get the following:

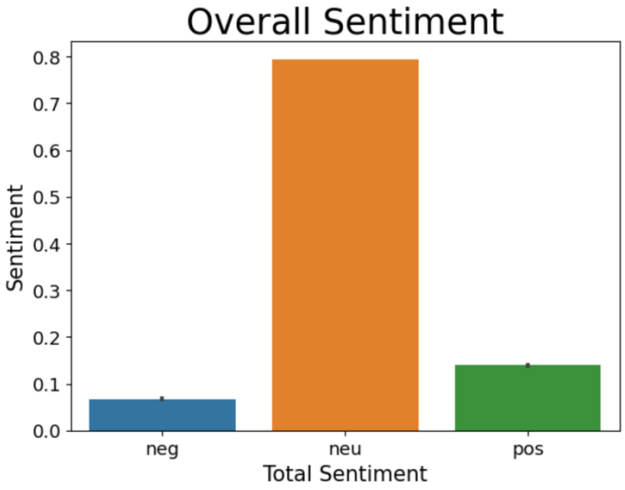
Train MAE: 43.28
Train RMSE: 44.
Train MSE: 2024.13

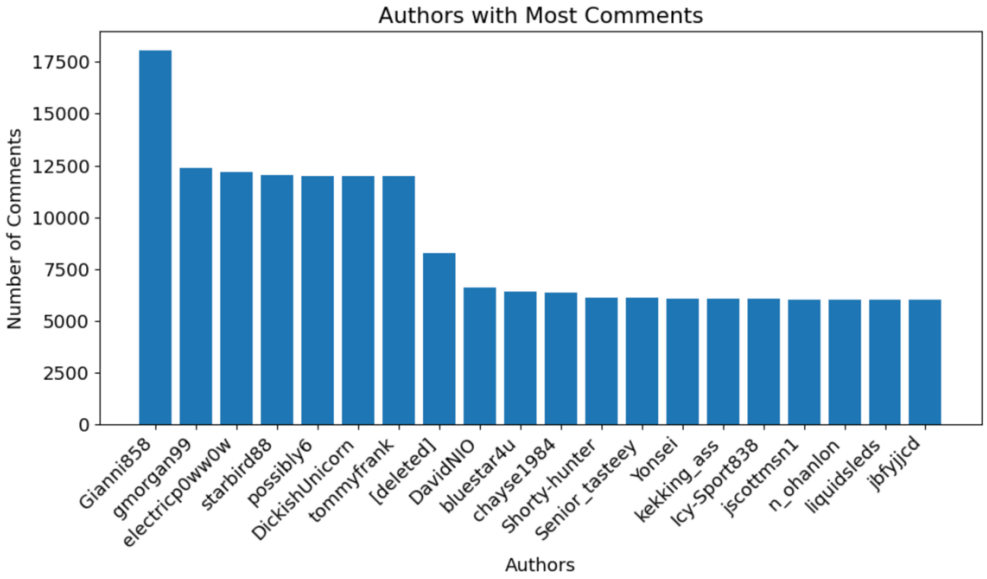
Test MAE: 43.66
Test RMSE: 44.24
Test MSE: 2024.13

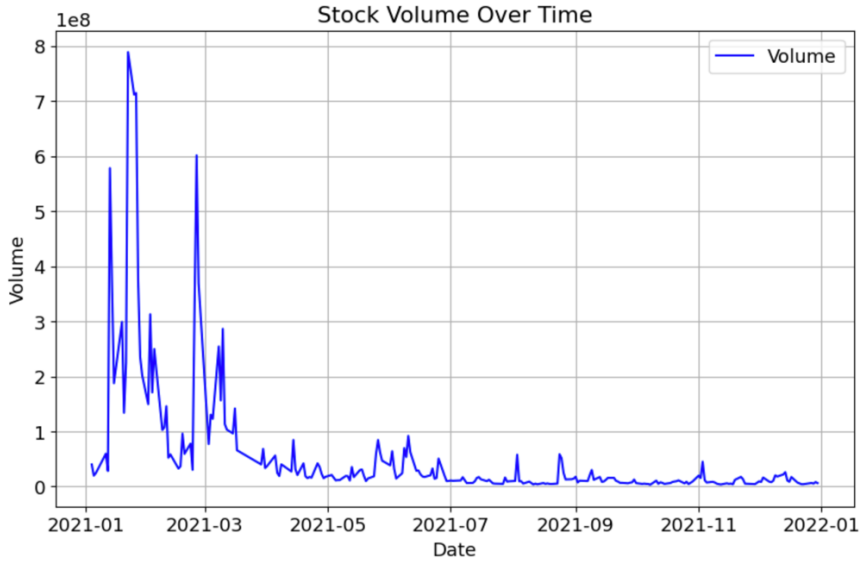
This doesn't make much sense, I believe my calculations were done incorrectly.

In addition to this, I looked into the sentiment analysis and of note I was able to get the most frequent words in the comments, an overall sentiment score, authors that commented the most, and volume each day.

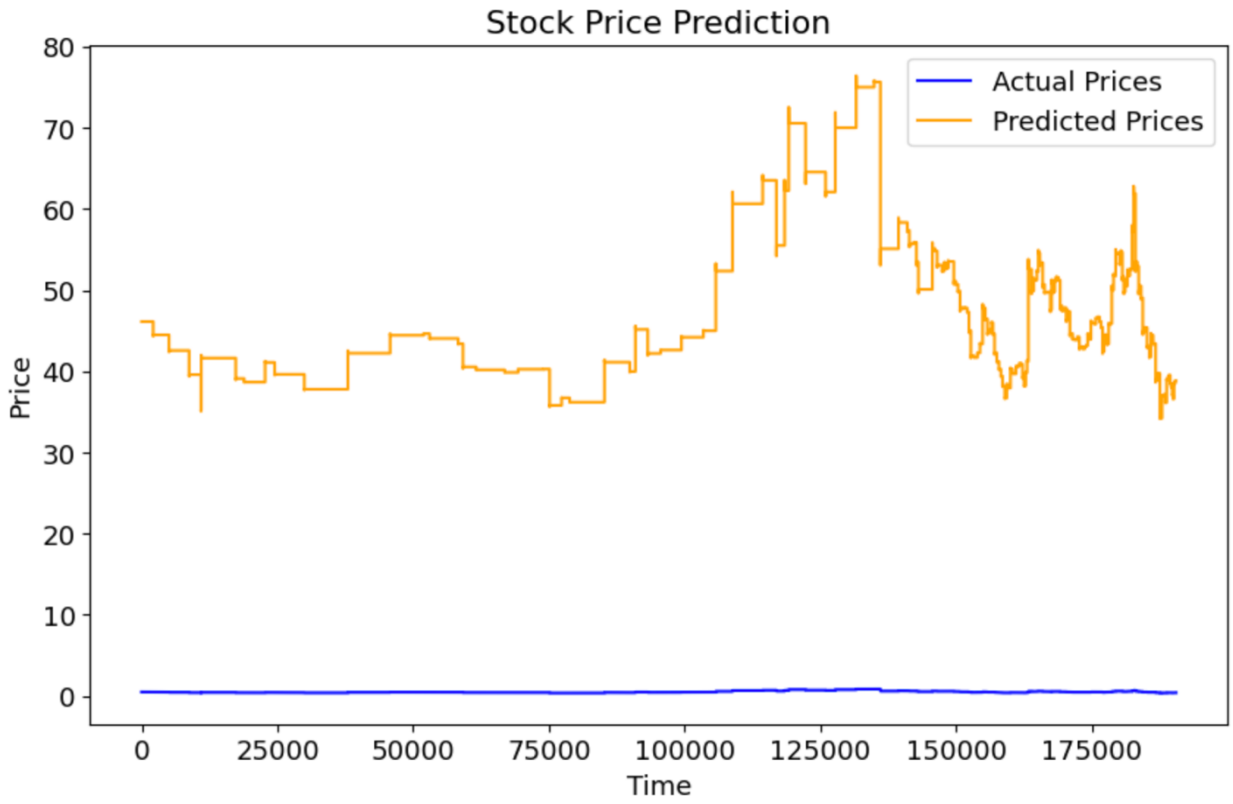

serious questions
fire sale
million dollars
strong hold
gme bae
finra gme
ape ireland
floor million
moon hope sound
gme data
best site
oh shit
rich fuck
moon hope
plausible scenario
routine still
Watch gladiator fellas
near future
hope sound
meaningful posts
moon audio
guys know
wants fire sale
diamond hands
hedgies get
wants fire
stimmy release


Overall Sentiment
Sentiment
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
neg
neu
pos
Total Sentiment


Authors with Most Comments
Number of Comments
0
2500
5000
7500
10000
12500
15000
17500
Gianni858
gmorgan99
electricp0ww0w
starbird88
possibly6
DickishUnicorn
tommyfrank
[deleted]
DavidNIO
bluestar4u
chayse1984
Shorty-hunter
Senior_tasteey
Yonsei
kekking_ass
Icy-Sport838
jscottmsn1
n_ohanlon
liquidsleds
jbfyjjcd
Authors


Stock Volume Over Time
1e8
Volume
0
1
2
3
4
5
6
7
8
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11
2022-01
Date

Lastly, I combined the LSTM model with the sentiment analysis data to identify any trends with the stock prices. To begin with, I use 'Close' as my feature as well as the compound, neg, neu, and pos sentiment scores. The LSTM model used has 50 hidden units (neurons) and uses the ReLU activation function. A dense fully connected layer predicts the next closing price with the adam optimizer. I use the mean squared error as the loss function, and I train it over 10 epochs. In the future this should be higher. I am not sure why my actual graph is showing a flat line, I plotted "y_test".



As for the performance metrics I get the following:

Test MAE: 436.27
Test RMSE: 47.22
Test MSE: 2229.47

I believe my model did well in terms of rising and falling fairly when there were sudden spikes in sentiments.

As I did not get to do much web scraping, I cannot comment on that process, however it has become harder and harder to scrap larger websites such as Twitter and Reddit as they incorporate anti-scraping methods. However, from an ethical issue, there are many issues of user privacy, bias and discrimination, and legal and regulatory challenges. Companies can only be proactive when it comes to the use cases of this data, but it ultimately the end user who must make wise decisions on what this data should be used for.

In conclusion, I learned a lot from this project and incorporated many of the learnings from the class. As the word embeddings, POS tagging, text stemming, lemmatization and broader cleaning of text was done for us, that made the project quicker to navigate. It was also interesting to see the emojis being converted into a vectorized form that aided in the development of the model. Going forward I would spend more time on hyperparameter tuning, and feature engineering. This would involve changing the batch size, optimizer and loss function. I would also automate the window size and see if that makes a significant difference in the LSTM. Of course, a better understanding of attention mechanisms can aide in the development of the models.

References:

- https://www.analyticsvidhya.com/blog/2022/01/the-complete-lstm-tutorial-with-implementation/
- https://regenerativetoday.com/implementation-of-simplernn-gru-and-lstm-models-in-keras-and-tensorflow-for-an-nlp-project/
- https://www.kaggle.com/code/michaelsammons/gamestop-eda-lstm
- https://github.com/tstewart161/Reddit_Sentiment_Trader
- https://wire.insiderfinance.io/game-stop-stock-market-price-prediction-using-rnn-lstm-7d6cfabfad51
- https://medium.com/swlh/is-social-media-mining-ethical-da4186d3b74b
- https://law.yale.edu/mfia/case-disclosed/social-media-mining-effects-big-data-age-social-media
- https://www.infosysbpm.com/blogs/web-social-analytics/data-privacy-and-ethical-considerations-in-web-and-social-media-analytics.html
- https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
- https://www.youtube.com/watch?v=ne-dpRdNReI