Stock Market Trend Prediction by News Headlines and LSTM

Yuzhou Liu
Department of Statistical Sciences
University of Toronto
yuzhou.liu@mail.utoronto.ca

Yingying Zhou
Faculty of Information
University of Toronto
yingying.zhou@mail.utoronto.ca

Abstract

Stock market investment is just another synonym for 'risk and return', the timely prediction of market direction remains one of the most challenging problems. Yet, in the end, there are always people behind investments and lots of decision making is based on what they read in the news. In order to provide guidance for individual investors, this paper proposes a deep learning-based stock index trend prediction model that considers the macro market sentiment and technical indicators. We develop an experimental framework for the classification problem which predicts next-day market direction. First, we incorporate the sentiment context and technical indicators in the baseline XGBoost model. Second, we adopt LSTM and further revise it by introducing ARIMA prediction as an extra feature. A novelty of this work is about the joint application of sentiment, technical indicators, and conventional time series techniques in the deep learning framework. Investors would be equipped with the toolkit of both fundamental analysis as well as technical metrics to make informative and profitable investments.

1 Introduction

The stock market is notoriously hard to predict. The stock index is a non-stationary financial time series characterized by volatility and noise such as investor irrational behavior and market sentiment, which are often difficult to be modeled by traditional techniques. Hypothesis Malkiel [2003] claims that the index is a 'random walk' ; the best guess for tomorrow's price would be the price of today. There is extensive time series research in the field of stock price study. However, the conventional analysis methods such as ARIMA Zhang et al. [2008] only model on the time series itself and impose strict statistical assumptions on data distribution and linearity, rendering the prediction result fairly off-target. To account for those limitations, the paper proposes to include context information as additional features in the stock index model and apply machine learning algorithms for the time series classification task. The remainder of the paper is constructed as follows. Section 2 reviews the previous work and highlights the novelty of our project; Section 3 describes the dataset; Section 4 presents the models and experiments; Section 5 delivers the result, followed by discussions and conclusion in Section 6.

2 Related work

Various machine learning tools along with featurization of technical indicators and investor sentiment have been used to study this 'random' financial time series.

2.1 Boosted tree model

Basak et al. [2019] built boosted tree models (random forest and XGBoost) to predict the direction of stock movements with technical indicators as features and achieved fair accuracy for 90-day time window prediction.

2.2 Long short-term memory model

Jin et al. [2019] predicted the individual stock closing price based on sentiment analysis and LSTM. The authors included 2 investor sentiment indexes (bullish/bearish), which are calculated through sentiment analysis on online stock comments, along with historical stock data into the price prediction model. The sentiment classification was performed by CNN with word2vec. Then, a Long Short-Term Memory (LSTM) model with attention mechanism was applied to predict the stock price of Apple. LSTM is a recurrent neural network with memory function, which is advantageous in analyzing time-series data. And as pointed out by Yan et al. [2018], the LSTM model on quantile regression outperforms conventional time series models that are commonly used in the finance field.

2.3 Novelty of our project

The existing literature inspires us to adopt a no-memory classification model as a baseline and continue to exploit the edge of LSTM.

The novelty of our work includes 1) we take a holistic market view where the national index (Dow Jones) is studied instead of individual stock prices; 2) we choose news headlines over investor sentiment to proxy for overall public sentiment; since the news coverage of macroeconomic conditions, political tensions and company activities are also influential factors to stock market performance; 3) we involve technical indicators in the LSTM model; 4) we use different sentiment analysis methodology (bootstrap subjectivity + polarity classification).

3 Data

3.1 Dataset description

Our working dataset is combined from time series stock data and sentiment context, which were retrieved from Yahoo Finance (`https://finance.yahoo.com/quote/5EDJI/history?=5EDJI`) and Reddit News (`https://www.reddit.com/r/worldnews/`) respectively. Data consists of 1,989 trading day observations from Aug. 8th, 2008 to July 1st, 2016. The stock data is daily Dow Jones index value including high, low, opening, closing, adjusted closing, and trading volume. The sentiment context is proxied by the daily top25 news headlines data. The 5-class sentiment scores of *Subjectivity*, *Objectivity*, *Positive*, *Negative*, and *Neutral* are derived from the news text data through a sentiment analysis algorithm and added into our dataset as new feature inputs. The label variable indicates whether the DJIA index (Figure 1) rose / fell for that day; '0' if the adjusted close value decreased, and '1' if the adjusted close rose or stayed the same. The dataset has a balanced number of labels, with 53.47% being '1' and the rest are zeros.

3.2 Feature engineering

Sentiment features The sentiment features are extracted from news titles using bootstrap classification for subjectivity and polarity borrowed from a Github repository Git [2017]; backed up by Volkova et al. [2013]. Each news title is decomposed into 5 classes of sentiments, with a probability score between 0% to 100%. As shown in the Figure 2, it is noticeable that most of the news headlines are not in a positive tone, (but not too negative either) and some of them are skewed towards subjectivity.

Technical indicators Technical indicators including 7-day moving average and exponential moving average were derived with stock index data and used as features for updated average index and weighted index changes.

ARIMA After testing the data from different periods, we found that each period represents a different pattern. Therefore, Auto ARIMA (AutoRegressive Integrated Moving Average) is used to predict daily DJIA close index value on a rolling basis in order to achieve the best fitting effect. The prediction result from ARIMA is added to the model as a feature in the LSTM models, in hope of denoising the series to some extent and extracting some new patterns.

To prevent machine learning leakage in the XGBoost model, closing-related data (close and adjusted close) are removed from the feature set. In addition, previous-day high and low data replace the same-day values. Eventually, the model is left with features *Volume*, *Open*, previous-day *high*, previous-day *low*, sentiment scores of 5 classes and technical indicators.

3.3 Data preprocessing

Days with incomplete news headlines have been filtered and the sentiment index with null values was imputed with corresponding column average. Sequential split was performed on the entire dataset in a proportion of 80 to 20 to form our training set (from 08 Aug. 2008 to 31 Dec. 2014, 1608 examples) and test set (from 02 Jan. 2015 to 01 July 2016, 378 examples).

4 Models

4.1 XGBoost classifier

Hyperparameters number of estimators is 25, maximum depth equals 3.

Features Lag-1 (previous-day) data for news sentiment scores, high, low, volume, closing values and technical indicators, and today's open index value to avoid data leakage and to involve as many dynamics as possible.

We start with a no-memory parallel boosted tree model, XGBoost, as the baseline model. XGBoost predicts a target by combining estimates from a set of weak classifiers so that error in classification successively decreases, and eventually delivers a strong result. Before feeding data into the model, we applied a Principal Component Analysis (PCA) to select the top 3 most important components through dimensionality-reduction. The intention here is to keep most of the significant information with little tradeoff of model accuracy loss. For the XGBoost classifier parameters, the number of trees is taken to be 1000, with the size of decision trees (depth) equal to 10.

4.2 Long short-term memory models

Features sentiment scores, historic DJIA index data, rolling ARIMA predictions, and technical indicators.

Data scaling Before running the LSTM models, the training set was scaled so that each data point has mean 0 and standard deviation 1. Standardization was also performed on the test set.

Timesteps LSTM models use sequences of 60 timesteps for the vanilla model and 180 timesteps for the revised model with rolling ARIMA. That is, we use data from the past 60 (or 180) days to predict today's stock market direction.

4.2.1 Vanilla LSTM

Long Short-Term Memory model is a Recurrent Neural Network (RNN) with memory function and forget gate so that overall meaningful information is extracted by the model from noisy data. For model hyperparameters, the activation function is ReLU, the number of hidden nodes is 128, the loss function is binary cross entropy, the learning rate equals 1e-5, with accuracy score to be the performance metric for this classification task.

4.2.2 LSTM with ARIMA as a feature

The rolling auto ARIMA predictions for daily index closing price is added to the LSTM model feature set in order to denoise the sequence and extract new patterns.

5 Results

5.1 XGBoost

The XGBoost has a classification accuracy of 56.1% on the test set with the help of principal eigen features identification from PCA. It is very computationally efficient and manages to detect some market patterns underneath the 'randomness'.

5.2 Vanilla LSTM

It turns out that the vanilla LSTM baseline model is not able to sufficiently incorporate important trend-relevant information from the noisy financial time series. The overall accuracy on the test set is 53.46%, which is better than a random guess (50/50 chance of success) by 7.0%.

5.2.1 LSTM with ARIMA as a feature

The introduction of ARIMA into the LSTM model significantly enhances the classification accuracy to 63.29% on the test set, surpassing the baseline prediction by 17.8%. The accuracy result indicates that the stock market is not completely unpredictable and that investors can harness the derived insights to beat the market.

6 Discussion

6.1 Results discussion

The model results in this paper are in line with literature findings that they outperform random guessing, therefore showing that the stock market is still 'predictable' to some extent. It is also verified that LSTM models can beat the baseline no-memory machine learning classifier (XGBoost). Notice that although our models are able to pick up some market direction signal, the accuracy is above 60%. To some extent, the result still reconciles with the semi-strong efficiency form from the long-standing Efficient Market Hypothesis (EMH) Jensen [1978] considering our forecast result is far from perfect. The hypothesis states that stock price reflects publicly-available information in a way that the market is almost 'efficient' so that technical and fundamental analysis are powerless in market forecasting.

6.2 Model performance comparison

As for why LSTM performed better than XGBoost, besides literature backup, the memory function of this recurrent neural network enables analysis of time series with order dependence between items in a sequence. Therefore, LSTM is more suitable for time series analysis than no-memory machine learning classifiers. We claim that the added feature of ARIMA prediction also plays a critical role in denoising sequence for LSTM models.

6.3 Limitations

There is still improvement space for LSTM models in capturing more signal information on market direction. Besides, the model is designed for next-day market trend prediction only. Such a short forecast window is not suitable for long-term investors and might pamper high-frequency short-term trading. Another problem would be the choice of sentiment proxy. As displayed in the exploratory analysis, news headlines tend to orient in a negative and subjective tone for publicity effects.

6.4 Future work

For further work, we plan to improve the LSTM with attention mechanism layers which might automatically identify and focus on impactful input components for market direction prediction. As for hyperparameter optimization, we would explore the option of reinforcement learning techniques in order to minimize classification loss. Different embedding methods can be experimented on sentiment analysis such as pre-trained BERT base, as an alternative for NLP emotion featurization.

Attributions

Group members contributed equally in this project and each of them participated in all tasks during researching, coding, and report writing with different focus. Yuzhou Liu was mainly responsible for code implementation and debugging. Yingying Zhou mostly focused on topic literature research, model tuning and modification, report writing, and workflow schedule. Both members devoted considerable time and energy to research and discussions. Throughout the whole process, they cooperated closely with each other.

References

Sentiment analysis source code. Github, 2017. https://github.com/ShreyamsJain/Stock-Price-Prediction-Model/blob/master/Sentence_Polarity/sentiment.py. Accessed: 2020-12-10.

S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, 47:552–567, 2019. URL https://doi.org/10.1016/j.najef.2018.06.013.

M. C. Jensen. Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, 6(2):95–101, 1978. URL https://doi.org/10.1016/0304-405x(78)90025-9.

Z. Jin, Y. Yang, and Y. Liu. Stock closing price prediction based on sentiment analysis and lstm. *Neural Computing and Applications*, 32(13):9713–9729, 2019. URL https://doi.org/10.1007/s00521-019-04504-2.

Burton G. Malkiel. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1):59–82, March 2003. doi: 10.1257/089533003321164958. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164958.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual Twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 505–510, 2013. URL https://www.aclweb.org/anthology/P13-2090.

Xing Yan, Weizhong Zhang, Lin Ma, Wei Liu, and Qi Wu. Parsimonious quantile regression of financial asset tail dynamics via sequential learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 1575–1585, 2018. URL https://proceedings.neurips.cc/paper/2018/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf.

D. Zhang, H. Song, and P. Chen. Stock market forecasting model based on a hybrid arma and support vector machines. In *International conference on management science and engineering*, pages 1312–1317. IEEE, 2008. doi: 10.1109/ICMSE.2008.4669077.

Appendices

A Tables & Figures

Table 1: Model Performance

		Comparison		
Model	Hyperparameters	Epoch	Sequence length	Accuracy
XGBoost	25 estimators, max_depth=3	1	all	56.08%
LSTM	128 hidden nodes, lr=1e-4	15	60	53.46%
LSTM_arima	128 hidden nodes, lr=1e-5	40	180	63.29%

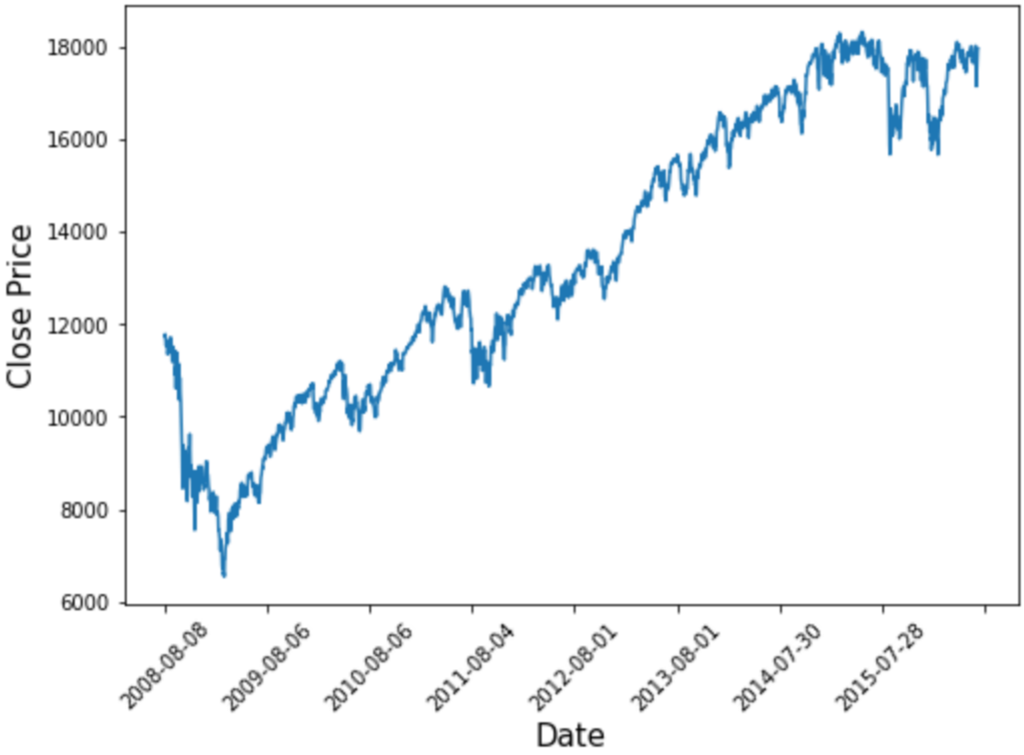

Figure 1: Dow Jones Index Adjusted Close from 08 Aug. 2008 to 01 July 2016

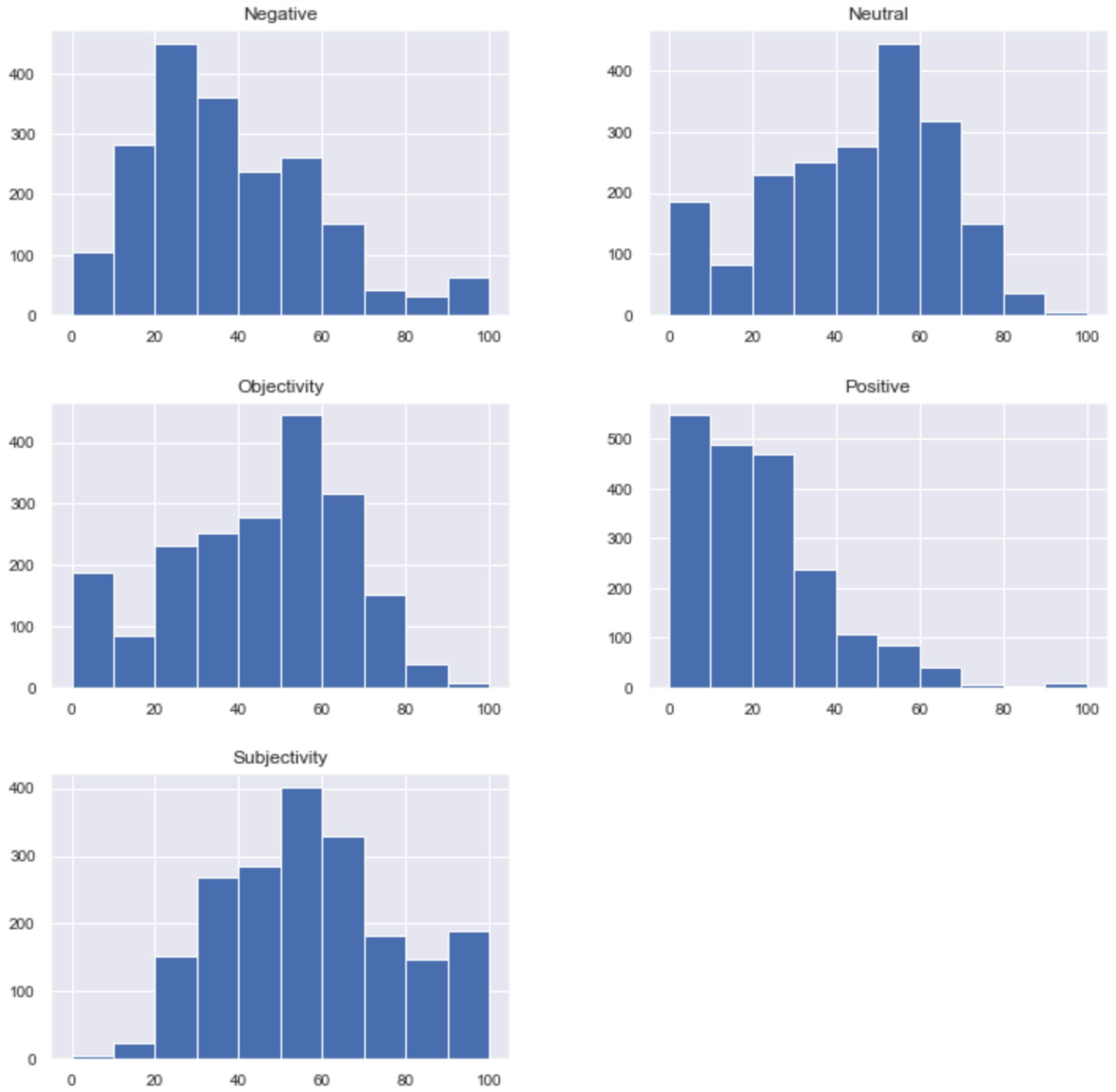

Figure 2: Sentiment Distribution of Top 25 News Headlines