**Literature Review for Seminar in Big Data**

* <https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2690?casa_token=Wss2Mc929EQAAAAA:mpS9-gJfV4oVku5GHmdFVBXTlmUgJvG1TvV2p7pOG0CXtFJ6LB4t_JBJEJEgZGgNMSx2zrTvQeq6zI8>
  + 2019
  + We use a sentiment index model, a kitchen sink logistic regression model, and a machine learning model (least absolute shrinkage and selection operator, LASSO) to forecast 1-month-ahead S&P 500 Index returns. In order to determine how successful each strategy is at forecasting the market direction, a “beta optimization” strategy is implemented. We find that the LASSO model outperforms the other models with consistently higher annual returns and lower monthly drawdowns.
* https://link.springer.com/content/pdf/10.1007/s00500-020-04862-3.pdf
  + 2020
  + a novel model named hybrid RNN model is proposed for stock market timing by incorporating multi-layer long short-term memory, multi-layer gated recurrent unit and one-layer ReLU layer. Moreover, based on five popular benchmark datasets from UCI Machine Learning Repository and six daily securities from Shanghai Stock Exchange, comparisons with 12 state-of-the-art models are conducted to verify the superiority of the proposed hybrid RNN model in terms of nine technical indicators. The findings from the experiment demonstrate that: (1) as opposed to 12 models, the average accuracy, MSE and AUC of hybrid RNN model (0.7406, 0.2592, 0.7368) are significantly better than other comparison models, and (2) the proposed hybrid RNN classification procedure can be considered as a feasible and effective tool for stock market timing.
* <https://www.mdpi.com/2571-5577/4/1/13/htm>
  + 2021
  + Use of NLP to predict stock prices from Stockwits Data (online platform)
  + We introduced FinALBERT, an ALBERT based model trained to handle financial domain text classification tasks by labelling Stocktwits text data based on stock price change. We collected Stocktwits data for over ten years for 25 different companies, including the major five FAANG (Facebook, Amazon, Apple, Netflix, Google). These datasets were **labelled with three labelling techniques based on stock price changes.** Our proposed model FinALBERT is fine-tuned with these labels to achieve optimal results. We experimented with the labelled dataset by training it on **traditional machine learning, BERT, and FinBERT models**, which helped us understand how these labels behaved with different model architectures. Our labelling method’s competitive advantage is that it can help analyse the historical data effectively, and the mathematical function can be easily customised to predict stock movement.
  + Labelling based on stock prices and not sentiment
* https://www.sciencedirect.com/science/article/pii/S1877050920307924
  + 2020
  + Artificial Neural Network and Random Forest techniques have been utilized for predicting the next day closing price for five companies belonging to different sectors of operation.