University of Warwick

Department of Computer Science

Research Proposal

Combining the Sentiment Analysis of News and Technical Indicators for Time Series Analysis of Stock Markets.

Author
Jack Wells

Supervisor
Dr. Tanaya Guha

February 20, 2020

Contents

1 Introduction

Stock market prediction is one of the most attractive research topics since the successful prediction can lead to significant profits, which has allowed for glamorisation as in movies such as “The Wolf of Wall Street” and “The Big Short”. Historically predictive methods are based on the analysis of historic market data, such as stock prices, moving averages and daily returns [12]. However, recent advantages in computing have allowed for the fast processing of a much more of the data, and so the context of the world around the company is now can now to be used to improve the modeling of stock prices.

Stock market prediction is considered one of the most challenging time series prediction due to the volatile nature and the noise the data contains [21]. Over the last few decades machine learning models, such as Support-Vector-Machines (SVM) and Support-Vector-Regression (SVR) have been used in the prediction of time series data, such as the financial data in this project [8] [13] [4] [18] [7]. These methods in these financial literature exploit sentiment signal features, which are inherently limited by not considering factors such as events and news context.

How to accurately predict stock price movement is still an open question, the Efficient-Market Hypothesis (EMH) [16] suggests that the stock price of a company reflects all available information and that any price changes are based on newly revealed relevant information, and thus the collection of this additional information will lead to more accurate models.

In this project it is hoped that the problem of not utilising all the data available will be addressed by leveraging deep neural models to extract rich semantic features from news text and from the opinion of stock brokers, it is hoped that this method is competitive with other state-of-the-art approaches demonstrating the effectiveness of Natural Language Processing (NLP) for computational finance. The use of NLP will hopefully be able to capture trends in data that are easily predictable given the information e.g. a change in the law or the private licensing of a product expiring may substantially effect a company.

2 The Problem

The simplest way to look at stock market prediction is the act of trying to determine the future value of a company [6]. Thus the problem in this project can be seen as a way of attempting to calculate the value of a company at a specific time, that may or may not reflect in the stock price of a company [21] for example it’s possible for a company to be worth more than the current stock price suggests if the future of that company is bright compared to the opposite in the case of a failing company. Using this information the future stock price should reflect the value of the company.

The problem this project intends to provide some insight into is the development of predictive models to accurately predict the future stock price of a company based on the information that is available, it is believed that the more contextual (textual data, brokers opinions) data that the model is given the better the predictions should be. However, it is fully possible that the increase in data doesn’t offer any advantage as the data may add noise, perhaps to an extent that no improvements can be made or that the extra data hiders the model.

3 Background

The area of stock market prediction is a well explored area going back many decades [6], attracting a considerable amount of research, with the methods changing over the years from using SVM to more modern methods like neural networks. The use of NLP in this area is a more recent development, with most of them focusing on a binary classification, a simplification of the problem of predicting a stocks price (classifying either an increase or decrease in stock price). Some of the most widely studied approaches rely solely on analysing the recent prices and volumes of the traded stock [19] [5] [14] [3] [11] and it is hoped that the model made in this project is competitive with these and out-performs them in many cases.

A model with the name Enalyst was introduced in Lavrenko et al and uses similar methods as this project intends to, that being the Yahoo finance to take the titles of news headlines. Vivek Sehgal et al. [20] and Michal Skuza et al. [15] used the sentiment analysis of texts to develop their models, with Michal Skuza paper using the analysis of tweets. However, all of these methods have limitations including unveiling the rules that may govern the dynamics of the market which makes the model incapable of capturing the impact of recent trends in the overall stock market.

Using the sentiment analysis of just textual data to predict an increase or decrease can't offer an accurate prediction on the actual price. Hence, the use of the sentiment analysis and stock market data, should be able to predict a future price of a stock.

Resent developments in the use of neural networks to learn dense representations of text, have been shown to be effective on a wide range of NLP problems, given enough training data. This provides strong grounds for the exploration of deep-learning based models in the prediction of stock prices. For example Ding et al. [23] has shown that the use of deep-learning representations of event structures yields better accuracy compared to discrete event features.

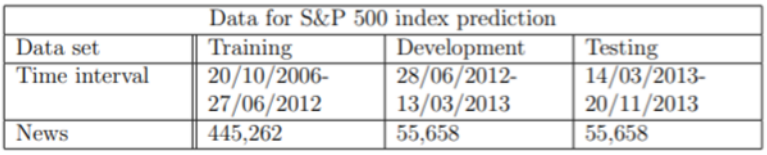
3.1 Key Literature

A key piece of work for this project is, “Leveraging Financial News for Stock Trend Prediction with AttentionBased Recurrent Neural Network” [12] by Huicheng, this project will develop on and use key insights on such as the use of newspaper headlines only instead of using more of the textual data for instance from the abstract of the article.

In this paper the author aims to leverage publicly released financial news and train an LSTM model to make predictions on the directional change of the both the Standard and Poor's 500 index as well as companies with in them. The model used by the author consists of Recurrent Neural Network (RNN) to encode the news text and capture the context information, self attention mechanism is applied to distribute attention on most relative words, news and days. The model outlined above is then demonstrated to show its competitiveness with other state-of-the-art approaches, demonstrating the effectiveness of recent advances in NLP technology for computational finance.

The paper then goes on to explain some of the key background in the methodology such as “Bag-of-Words” and “LSTM” models. This is something that will be done in detail in the proposed project but not explained with in it.

The data that the model in [12] was developed on was collected from Reuters and Bloomberg and the data has been made publicly available by Ding et al. [22] a summary can be seen below. This data will make an excellent start to building a model that can be used in the proposed project, as collecting and manually classifying data will be a challenge to do on the scale that is required to make an effective model (this is discussed further in the Risks section). Hence, why this project aims to use this same source of data to train the model on and then fine tune the model with more relevant data from more recent FTSE 100 data. As has been done in Huicheng's paper.

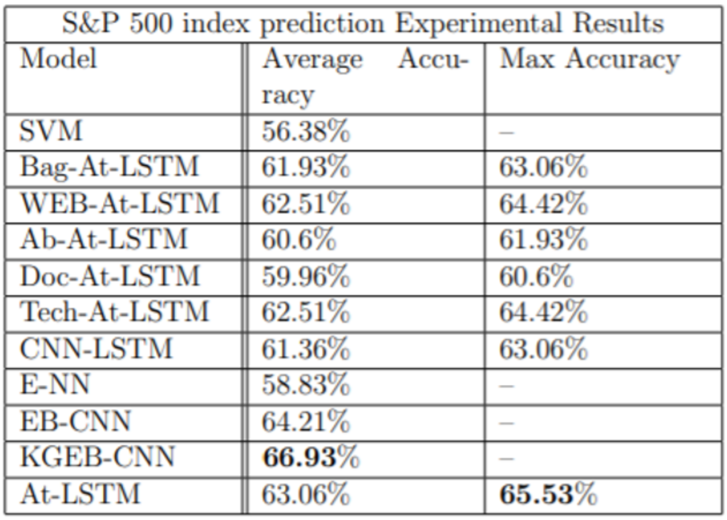


Data for S&P 500 index prediction			
Data set	Training	Development	Testing
Time interval	20/10/2006- 27/06/2012	28/06/2012- 13/03/2013	14/03/2013- 20/11/2013
News	445,262	55,658	55,658

Figure 1: Data for S&P 500 index prediction

The paper uses a pre-trained 100 dimensional word embedding with a skip-gram method, trained and developed on the table in figure 1, with a training vocabulary size of $153,214$ words. The word embedding are fine-tuned during the model training.

The results of this paper are shown below for several different models. Those being an SVM a form of classical machine learning, the others are forms of deep neural networks. The Bag-At-LSTM and At-LSTM show that sentence encoding with LSTM models offers a slight accuracy improvement than a Bag-of-Words model. The WEB-LSTM and At-LSTM indicates that character level composition helps improve the models accuracy. The comparison between CNN-LSTM and At-LSTM shows that the news level self-attention layer can help capture more relevant news titles and their temporal features. The author reaches the same conclusion as Ding et al. [22], that the use of news headlines offers a higher accuracy than using the whole article or just its abstract, both of which have some negative effect on the model. As shown in the Ab-At-LSTM and Doc-At-LSTM. The model proposed by the author offers a lower accuracy than the accuracy of KGEB-CNN proposed in [22]. As discussed by Huicheng [12] this is likely because Knowledge Graph Event Embedding (KGEB) is a more powerful method than the sequence embedding preformed in Huicheng's paper.

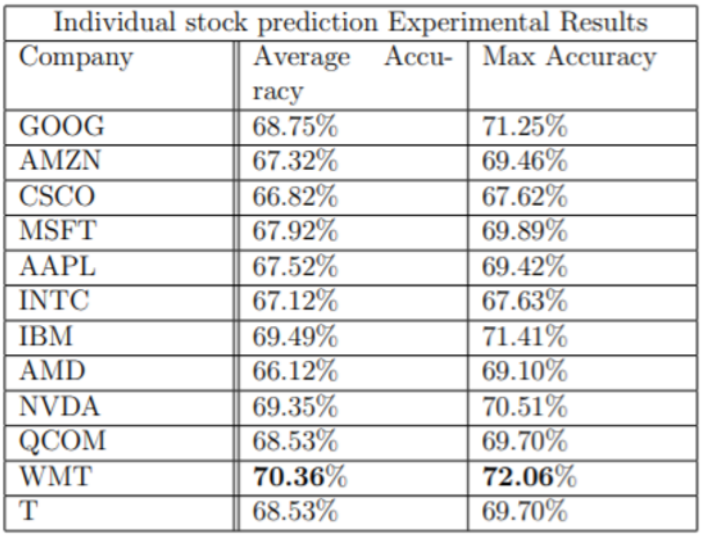


S&P 500 index prediction Experimental Results		
Model	Average Accuracy	Max Accuracy
SVM	56.38%	–
Bag-At-LSTM	61.93%	63.06%
WEB-At-LSTM	62.51%	64.42%
Ab-At-LSTM	60.6%	61.93%
Doc-At-LSTM	59.96%	60.6%
Tech-At-LSTM	62.51%	64.42%
CNN-LSTM	61.36%	63.06%
E-NN	58.83%	–
EB-CNN	64.21%	–
KGEB-CNN	**66.93%**	–
At-LSTM	63.06%	**65.53%**

Figure 2: Results showing the accuracy of using different types of neural networks with different forms of data and data pre-processing.

As can be seen the SVM model performs significantly worse than the neural network approaches with the accuracy of the models made by Huicheng average about 62% accuracy that is over 5% better than the SVM model.

There however are more promising results shown when the analysis is done on an individual company (figure 3), this is likely due to the news articles being used to predict being more relevant. In contrast to using the entire corpus to predict the s&P 500 which adds noise and thus reduces accuracy. The accuracy of predicting WALMART is over 70% which is better than the best model of predicting the entire S&P 500.



Individual stock prediction Experimental Results		
Company	Average Accuracy	Max Accuracy
GOOG	68.75%	71.25%
AMZN	67.32%	69.46%
CSCO	66.82%	67.62%
MSFT	67.92%	69.89%
AAPL	67.52%	69.42%
INTC	67.12%	67.63%
IBM	69.49%	71.41%
AMD	66.12%	69.10%
NVDA	69.35%	70.51%
QCOM	68.53%	69.70%
WMT	**70.36%**	**72.06%**
T	68.53%	69.70%

Figure 3: Results of the accuracy of using the At-LSTM model on predicting individual companies.

Another key piece of work is "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation" [22]. This paper explains how the use of Open Information Extraction (Open IE) have enabled the extraction of structured events from web-scaled data. This method is shown to offer significant improvements in the accuracy over the use of a bag-of-words approach in the sentiment analysis. This paper mainly contains information on how to effectively use these NLP methods for the prediction of stock prices. The model uses an additional evaluation metric based on the Matthews Correlation Coefficient (MCC) to avoid bias due to data skew. This is because some companies may have a large number of increases in there stock price, this data would have a high accuracy even if the predictive model classified all points as positive. This paper is particularly useful as it develops a way of processing financial textual data and makes available the large amount of data that they used which will be used in the training of the model in this project. See figure 1 for details of the data set.

These previous pieces of related work offer many key insights into the problem of stock price prediction, and offer many approaches that can be implemented in this project.

4 Project Aims

This project aims to develop a model to predict the future stock price of a company the metrics that will be use to develop measure the accuracy of the model will be the Mean

Squared Error (MSE) which is a measure of how far the prediction is compared to the actual value. A variety of different methods will be used to make the models including the use of classical machine learning techniques such as support vector machines and logistic regression as used in [12]. A LSTM model will also be developed that is aimed to make predictions better as has been done in [24]. This is because deep learning techniques have been shown to offer improved accuracy. Other metrics will be used to compare the model in this project against other models such as those in the key literature section, this is discussed more in metrics and results section.

5 Methodology

This section outlines how a solution will be implemented and details the methods that will be used and information on how the data will be collected.

LSTM models are used as this type of neural network aims to keep a "memory" of previous events, it was first proposed by Hochreiter in 1997 [17]. LSTM units are a building unit for layers of a Recurrent Neural Network (RNN), a RNN composed of LSTM units are often called an LSTM network. LSTM models are able to learn long term dependencies, this makes them an ideal form of RNN for the classification, process and predicting time series given time lags of unknown size and duration between important events. LSTM's also offer promising results in the sentence encoding in many NLP applications [10] [9].

5.1 Data Source

The data that will be used is taken from Yahoo finance [2], the project will implement three types of web scrapers one that will take stock market information such as the volume traded, opening stock price, closing stock price, etc. and produce CSV files containing that data over a given time period. This is available from the Python "Pandas" library that contains a method for collecting this data. This project will also extract the most recent news articles (going back one month) concerning a given company and use this for text analysis. The title of the news articles will be used exclusively as this has been shown to give more reliable results [22] than using the entire article or just the abstract, this is most probably due to the extra noise that the data in the full article. It also makes the collection and data processing easier, which will allow for faster training of models.

However, the collection of enough data to produce a reliable sentiment analysis model isn't feasible so this project will make use of the data set was made publicly available by Ding et al. [22], as shown in the related work section.

This project also aims to collect data on from the opinion of specialists in the area of stock price prediction. The opinion of brokers and their future prediction will be obtained, these brokers will classify a stock into categories such as: "Buy", "Hold", "Sell", "Non-performing", "Overweight" and "Top pick". The brokers opinions and the predicted future price from the brokers should be good predictors for the longer term stock price of a company. This data can again be collected by the use of a web scraper used on the "equiniti" website. [1]

The use of web scrapers as described above allow for the easy and quick collection of data, it also allows for an easy way to extend the data. For instance training on the FTSE 250 or other stock markets such as the S&P 500, as they only need the stocks code (e.g. TSCO for Tesco PLC) to collect all the data needed.

5.2 Process

This project will require the building of two models for prediction, those being one to classify the sentiment of a headline and the other to take all the data including the sentiment to produce an expected value of a companies stock price.

For the first model to predict the sentiment of the text, this project will look into several different methods of doing this such as the use of a bag-of-words method and the vectorization of the words. This is where words are represented by a vector and the angle between the words represents how similar the words are, i.e. “good” and “great” have a smaller angle between them than “good” and “bad”. Where as the bag-of-words method counts the frequency of words used in the articles of each category. This does mean that the title of an article will have to be categorised by sentiment (e.g. into “Positive”, “Neutral” and “Negative”). This model can then be a classical model such as a Naive Bayes or a neural network, the choice of model will be selected through experimentation to test there accuracy. The hope being that the neural network will perform better and that the classical model will be used as a bench mark. One of the key advantages of using a neural network to classify the sentiment is that one of the last layer of the model can be used as an input in the second predictive model other than a categorical prediction, this can hopefully offer a greater insight into the sentiment than a simple categorical classification.

The second model will be used to predict the value of a companies stock price taking input from the first model as well as the stock price data and later the broker opinion data. In this project a classical method will be implemented such as an SVM as done in previous works to bench mark the performance of an LSTM model [13] [8]. As in other works it is expected that the LSTM model will out perform the classical methods.

The increase in run time for the neural network models is not an issue as most of the data is published with a low frequency and hence constantly updated predictions aren’t necessary.

5.3 Metrics and Results

The aim of this project is to accurately predict the future price of a company hence a good metric for this would be the Mean Squared Distance (MSD) between the actual value and the predicted value. However, for comparison to other works the predicted value can be used to create a binary classification the same way as was done in [12], to do this simply use the predicted value and classify the predicted value into two categories (possibly more) those being “Increase” and “Decrease” based on if the price has increased or decreased since the day before, this will allow me to make direct comparison between the model in this project and models made in other related works such as [22] and [12]. Using a categorical metric will also show if any particular events e.g. the price increasing is being over predicted by using a confusion matrix. It also allows for the use of the Matthews Correlation Coefficient (MCC) which avoids bias due to data skew, and is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where TP, FP, TN and FN are True Positive, False Positive, True Negative and False Negative, respectively.

The results of the different models will be easily shown in a table format which will allow for comparison between them to see what combination produces the best predictive model.

6 Project management

6.1 Time Management

The time-line for this project ranges from 01/01/2020 to 01/09/2020. Various stages have been set out in order to monitor the progress in this project and allow for the right level depth in this project as to keep up with the time-line. For instance there is no need to develop a LSTM model with many layers if doing so will offer little benefit to the accuracy and may cause time pressure on more important parts of the project such as the NLP modeling. Please see the table in the appendix to see a break down of the stages required.

During stage 1, a simple model is going to be produced based solely on previous stock prices, and then developing it with the implementation of NLP and if time permits the use of the brokers opinion data. Later stages will aim to fine tune and add all available data to hopefully allow comparisons between the models which use the different data.

6.2 Risk Management

A main risk in this project as mentioned earlier is the collection of enough data to effectively train a NLP neural network on. As has been done in other papers similar to this project the use of the publicly available data set by Ding et al. [22] will be used. However, as the data is collected on companies based in the United States it may not be compatible with UK stocks to mitigate this the model will be fine tuned with recent news headlines collected on the FTSE 100. This will reduce the problem however, the problem of time it takes to manually classify these tuning data news titles will still be an issue.

Due to the nature of this problem the dependency of the future price may vary for individual companies and thus prevent generalisation for the period of time required to predict the next days stock price on.

These problems culminate into another major challenge in this project that being the successful implementation of a model. Such that known results are reproduced with similar results. Due to the type of data there may be large amounts of noise in the data that may make it hard to predict the future and using less data may offer better results. This has been shown to be the case in [12] where the use of the full article produced lower accuracy compared with using just the headline of the news article.

6.3 Current Progress

At this time in the project the majority of the work has been focused on reading and exploring the literature around this subject in an effort to formulate a successful way to solve the problem by following the lead of previous works in this area. The development of models has just begun, currently testing these models on the stock price data only to check the feasibility and gain an understanding of how to implement the LSTM models in Python. A large proportion of the coding has been given to the creation of the web scrappers as to build up a reserves of data, as some of it can't be accessed all the time e.g. the news headlines only go back one month on Yahoo Finance. Currently all the data collection tools have been created and used to save the relevant information for future use.

6.4 Ethics and Intellectual Property

The use of web scrapers does impact the owner of a websites as making many requests to a server in a couple of seconds will reduce the websites ability to serve normal website traffic. However, using Yahoo as a main source of data their infrastructure is capable of handling this many requests with minimal impact, also as news articles aren't published as frequently. This means that request for the news articles can be limited to once or twice a day. In regards to the use of equiniti website again the number of requests can be limited to a relatively small number per day (approximately 100). This should hopefully not impact the owners of the websites too much. From a legal point of view as they data collected from the websites isn't being used maliciously and is for research purposes there should no reasons that this project should cause concern to anyone.

This project will not specifically develop any new or novel methods or procedures which intellectual property protection would be required. The copyright of the written material such as in the news headlines will remain with the author. However, free and open access to all material is granted under the creative commons licensing scheme to Warwick University staff and students for the purposes of further research activity.

7 Conclusion

Existing research shows that the use of textual analysis in the prediction of stock prices does offer an improved predictive model and that the use of LSTM models with and without the use of textual data also offers an improvement compared to the use of classical machine learning methods such as the SVM. [12]. However, it appears that there has been less research into the prediction of an exact price and research in the use of combining more than one form of extra information such as using text analysis and brokers opinions is lacking. This project aims to address this and compare the different forms of data and how they interact in the prediction of stock prices.

7.1 Developments

Given time constraints allow there are many different developments that can be made these include some of the previously discussed additions such as using data from other stock markets to make a more generalised model or even making a second model trained on that data exclusively and testing between them to see if the models are transferable.

References

[1] Equiniti website for brokers opinions, https://equiniti.moneyam.com/broker-views/.

[2] Yahoo finance, https://uk.finance.yahoo.com/.

[3] C. K. Ayo A. A. Adebiyi, A. O. Adewumi. Comparison of arima and artificial neural networks models for stock price prediction, 2014.

[4] G. Francis A. N. Refenes, A. Zapranis. Stock performance modeling using neural networks: a comparative study with regression models, neural networks, 1994.

[5] V. Akgiray. Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts, 1989.

[6] K. McCue B. G. Malkiel. A random walk down wall street, 1985.

[7] C.-C. Chiu C.-J. Lu, T.-S. Lee. Financial time series forecasting using independent component analysis and support vector regression, decision support systems, 2009.

[8] S. Pal D. Basak and D. C. Patranabis. Support vector regression, 2017.

[9] J. Schmidhuber F. A. Gers, N. N. Schraudolph. Learning precise timing with lstm recurrent networks, Aug 2002.

[10] Q. V. Le I. Sutskever, O. Vinyals. Sequence to sequence learning with neural networks, 2014.

[11] H. Ahn K.-J. Kim. Simultaneous optimization of artificial neural networks for financial forecasting, 2012.

[12] Liu and Huicheng. Leveraging financial news for stock trend prediction with attention-based recurrent neural network, Nov 2018.

[13] E. Osuna J. Platt B. Scholkopf M. A. Hearst, S. T. Dumais. Support vector machines, 1998.

[14] A. Boru A. T. Dosdogru M. Gocken, M. Oz¸calıcı. Integrating meta-heuristics and artificial neural networks for improved stock price prediction, 2016.

[15] A. Romanowski M. Skuza. Sentiment analysis of twitter data within big data distributed environment for stock prediction, 2015.

[16] B. G. Malkiel. The efficient market hypothesis and its critics, 2003.

[17] J. Schmidhuber S. Hochreiter. Long short-term memory, 1997.

[18] S. Padhy S. P. Das. Support vector machines for prediction of futures prices in indian stock market.

[19] E. Schoneburg. Stock price prediction using neural networks: A project report, 1990.

[20] C. Song V. Sehgal. Sops: stock prediction using web sentiment, 2007.

[21] Baohua Wang, Hejiao Huang, and Xiaolong Wang. A novel text mining approach to financial time series forecasting, Dec 2011.

[22] T. Liu J. Duan X. Ding, Y. Zhang. Using structured events to predict stock price movement, 2014.

[23] T. Liu J. Duan X. Ding, Y. Zhang. Deep learning for event-driven stock prediction., 2015.

[24] S. K. Halgamuge Y. Zhai, A. Hsu. Combining news and technical indicators in daily stock price trends prediction, Aug 2007.

8 Appendix

8.1 Time Management Table

Description		Action
Stage 1	Initial Research Phase (February-April)	In this stage an exploration into the current research and how the research has been implemented will take place with the development of simple models.
Stage 2	Skills Development (Ongoing)	The skills needed for this project will be developments beyond the scope of many modules such as NLP, Data Mining and Data Analysis. This project will also require a deep understanding of neural networks and there implementations in Python. All of which will be learned throughout the duration of the project.
Stage 3	Presentation (March-April)	Preparation and presentation of the research at the current stage of the project. This will most likely be a general over view of the topic and some of the research in this area.
Stage 4	Exam Period (March-May)	During this period minimal development of the project will tale place as revision for the end of year exams will be take priority.
Stage 5	Implementation (May-June)	During this stage development of a model will be finalised and the hyper-parameters tuned. The model will also be developed by changing the number of layers in the LSTM model to make a models with the best accuracy.
Stage 6	Results (July)	During this stage the models will be used to make predictions, the model will be used to predict the price, as described above if the price increases or decreases the results looking at the data this way can be used to compare with the related work and the MSD can be used to calculate how good the model is.
Stage 7	Write-up (August-September)	This stage will be used to compile all of the previous stages into one document containing all the relevant information that this project has used.