Do News and Sentiment play a role in Stock Price Prediction?

Vanstone, Bruce J; Gepp, Adrian; Harris, Geoffrey

Published in:
Applied Intelligence

DOI:
10.1007/s10489-019-01458-9

Do News and Sentiment play a role in Stock Price Prediction?

Abstract. Despite continuous improvement in the range and quality of machine learning techniques, accurately predicting stock prices still remains as elusive as ever. We approach this problem using a modern autoregressive neural network architecture and incorporate sentiment predictors, which are becoming increasingly available due to advances in text mining techniques. We find that the inclusion of predictors based on counts of the number of news articles and twitter posts can significantly improve the quality of stock price predictions.

Keywords: Stock prices, Sentiment, Auto Regressive Neural Networks, News, Twitter

1 Introduction

This paper extends a previously published conference paper (Vanstone, Gepp, and Harris 2018) detailing our initial investigations into the use of sentiment metrics to enhance the stock price prediction process.

Predicting stock prices is a difficult problem, as stock prices are inherently noisy observations of random variables, which in turn represent the sum of investors' future expectations about a company's value. There have been many attempts to use machine learning techniques to aid in the stock price prediction problem. Most academic work in this area can be broadly classified as either econometric modelling (Rapach and Zhou 2013), or machine learning (Atsalakis and Valavanis 2009). The fields of financial investment and trading are dominated by the fundamental analysis framework, and the technical analysis framework respectively. Each of the financial frameworks provides for the calculation of variables that can be used to assess the financial state of a company from an investment perspective.

Technical Analysis relies on exchange generated data, specifically price and volume data at specific points in time. As the set of variables available from exchanges is quite small, this implies that the majority of technical variables are essentially different mathematical derivations of the same underlying price or volume data, over varying timeframes. For this reason, machine learning methods using many technical variables are essentially just increasing the noise in the modelling process by incorporating increasing amounts of covarying data.

On the other hand, fundamental variables may have the potential to offer additional information, however, their disclosure is usually annually or semi-annually, and as such, they are not available in the required frequency for shorter term price prediction.

The Efficient Market Hypothesis (EMH) (Fama 1965) is the primary theory in finance relevant to price prediction. This hypothesis asserts that a stock price instantaneously reflects all available information implying that prices react instantaneously to news and it should not be possible to outperform the market. It should be noted that the extent to which markets are considered efficient is somewhat controversial (Yen and Lee 2008).

As financial theory points to the origin of stock price changes being a response to new information, and as market prices represent the combined views of investors' expectations of a company's future value, there is every reason to expect that news articles and twitter opinions may represent exogenous variables which may be useful in shorter-term price prediction.

Theoretical models of the effect of investor sentiment usually posit the existence of two types of traders: 'noise traders', who hold random beliefs about future dividends, and 'rational arbitrageurs', who hold Bayesian beliefs (Long et al. 1990). It is reasonable to assume that noise traders may be influenced by negative news stories, which would lead them to sell investments to rational arbitrageurs, creating temporary downward pressure on prices.

In 2007, using text from the Wall Street Journal's Abreast of the Market column, Tetlock (Tetlock 2007) finds that high media pessimism predicts short-term downward pressure on market prices followed by a reversion to fundamentals, and further that unusually high or low pessimism predicts high market trading volume. These results are consistent with the theoretical finance models and are inconsistent with the theory of media content as a proxy for new information about fundamental asset values. This is because if media content was a proxy for new information, then there would be no expectation that market prices would revert back to prior fundamental values, instead, the new information should establish a new fundamental value.

It appears then, that noise traders may well sell stock to rational arbitrageurs after periods of negative news. Rational arbitrageurs exploit the temporary drop in stock prices to acquire stock with the expectation that it quickly returns to fundamental value, thus obtaining a profit. As the primary drivers of this approach are the rational arbitrageurs, we source our news and sentiment scores from Bloomberg, which ensures we use a source applicable to sophisticated investors.

In this paper, for each of Australia's 20 largest stocks, we build two Neural Network Autoregressive (NNAR) models: one a basic NNAR model, and the other an NNAR model extended with counts of news articles and twitter posts. By comparing the prediction accuracy of the two models, we aim to assess whether the inclusion of sentiment variables based on the count of news articles and twitter posts can enhance the accuracy of the stock price prediction process.

2 Methodology

2.1 Sentiment Scores

This paper sources sentiment scores from Bloomberg that are based on both news articles and Twitter. Sentiment scores and closing price data are obtained for all stocks in the S&P ASX20 from the start of January 2015 to the end of July 2018. The starting date is the date that Bloomberg first began publishing the sentiment scores.

Bloomberg publish six sentiment scores on a daily basis. Table 1 shows the Bloomberg field names and brief description of the six sentiment score variables.

News and Twitter sentiment scores are calculated by Bloomberg using an undisclosed proprietary approach. As such, the exact methodology used to calculate

the sentiment scores is not disclosed. However, this is not an issue for this work, as we are specifically interested in how future prices are influenced by the reactions of professional investors to the published scores. Further, after feature reduction (see Section 2.2), the important variables for both the news and twitter data were revealed to be only the count variables. As such, the mechanics of the sentiment calculations are unimportant in this work.

Although the positive and negative sentiment variables are not used in this paper, for completeness, we provide a quick precis of the available information relating to the Bloomberg methodology. In Bloomberg's methodology, a human expert initially reads each article and assigns a positive, negative, or neutral label, based on their judgement of how the news story would affect an existing holder of that security. These manually created cases are then provided to a machine learning model which, when trained, is capable of taking a news story and providing a probability of whether the story could be expected to have a positive, negative, or neutral effect on a pre-existing holder.

In practice, news is released in continuous time. For each news article, Bloomberg determines article-level sentiment, consisting of both a score and confidence. The score value is 1, 0, or -1, representing positive, neutral, or negative sentiment predictions from the model. The confidence ranges from 0 to 100, representing a likelihood probability. Bloomberg then produce company-level daily sentiment scores, which are a confidence-weighted average of the past 24 hours of story-level sentiments. These company-level scores are published every morning approximately 10 minutes before the market opens.

Bloomberg Field Name	Brief Description	Name used in this paper
NW043	News publication count	NEWS COUNT
NW044	News positive sentiment count	NEWS POS
NW045	News negative sentiment count	NEWS NEG
NW039	Twitter publication count	TWITTER COUNT
NW040	Twitter positive sentiment count	TWITTER POS
NW041	Twitter negative sentiment count	TWITTER NEG

Table 1: Bloomberg Sentiment Data

2.2 Feature Selection

When the goal of modelling is prediction, it is important to select the minimum set of attributes that are required to build the model. Feature reduction is key to building parsimonious models, as unnecessary features add noise but do not necessarily increase predictive ability. In an effort to build as simple a model as possible, feature selection was performed amongst each of the sets of three Bloomberg fields describing the news and twitter sentiment. In this paper, a data-driven feature selection process was performed by fitting linear models to the in-sample data, and selecting only the feature with the highest variable importance.

Firstly, a linear model was fit to the in-sample data using the Bloomberg fields NW043, NW044 and NW045, with the objective of determining the most important News variable. The fitted model was then used to extract individual

variable importance. Variable importance for linear models is calculated from the absolute value of the t-statistic, and scaled to have a maximum value of 100. The variable importance amongst the News variables is shown in Table 2.

Bloomberg Field Name	Brief Description	Relative Importance
NW043	News publication count	100.00
NW044	News positive sentiment count	34.38
NW045	News negative sentiment count	0.00

Table 2: Variable Importance amongst News variables

A linear model was then fit to the in-sample data using the Bloomberg fields NW039, NW040 and NW041, with the objective of determining the most important Twitter variable. The variable importance amongst the Twitter variables is shown in Table 3.

Bloomberg Field Name	Brief Description	Relative Importance
NW039	Twitter publication count	100.00
NW040	Twitter positive sentiment count	38.71
NW041	Twitter negative sentiment count	0.00

Table 3:Variable Importance amongst Twitter variables

As shown in Table 2 and Table 3, the publication count variables were the most important, and the variables which were used for further modelling were chosen as NW043 and NW039.

Initially, it may seem somewhat counter intuitive that the total publication counts were chosen over the positive and negative counts for both News and Twitter. We hypothesize that this suggests that investors are sensitive primarily to the fact that there *is* news about their chosen investments, rather than the nature of that news.

2.3 Neural Network Autoregressive Models

In this paper, we fit Neural Network Autoregressive NNAR(p,P,k) models. These are feed-forward, single hidden layer neural networks that use lagged inputs, and are well suited to forecasting non-linear univariate time series data.

For non-seasonal time series, the fitted model is denoted NNAR(p,k), where p is the optimal number of lags (according to AIC) for a linear AR(p) model, and k is the number of hidden nodes. This is analogous to an AR(p) model but with nonlinear functions. For seasonal time series, the fitted model is denoted NNAR(p,P,k)m, which is analogous to an ARIMA (p,0,0)(P,0,0)m model, but with nonlinear functions (where k is still the number of hidden nodes). In the case of seasonal time series, the defaults are $P = 1$ and p is chosen from the optimal linear model fitted to the seasonally adjusted data (Hyndman and Athanasopoulos 2013).

Two NNAR models are created for each stock. For each model built, a total of 20 networks are fitted, each with random starting weights. These are then

averaged when computing forecasts. The networks are trained for one-step forecasting, and multi-step forecasts are computed recursively. The number of hidden nodes in each network is half of the number of input nodes (including external regressors) plus 1.

2.4 Neural Network Autoregressive Models

Initially, the data is split into two parts: all the data except the last month is used to train the NNAR models and the final month is used for out-of-sample testing. The in-sample training data thus covers the period 01-01-2015 to 31-05-2018. The out of sample data covers the period 01-06-2018 to 30-06-2018.
To assess the predictive value of the sentiment scores, two NNAR models are built for each stock using the in-sample data. The first model (BASIC) uses only lagged values of price to predict future prices. The second model (SENTIMENT) extends the first model by supplying the News publication Count (NW043) and the Twitter Publication Count (NW039) as additional predictors. The future (out-of-sample) predictions from each of the models is then compared to the actual future prices, and an RMSE is calculated for each model. We then compare the two RMSEs for each stock to determine which model, BASIC or SENTIMENT, was the closest fit to the actual future prices.

Figure 1 shows the workflow applied to each stock in the S&P ASX20. All code was developed in R, version 3.3.3.

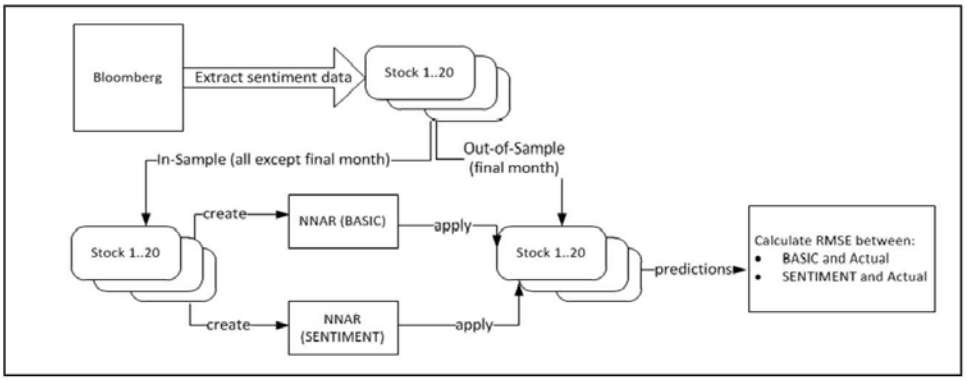

Figure 1: Workflow

3 Results

Table 2 shows the summary statistics for the downloaded stock data. For each stock, there are three rows showing the number of observations and the mean of each of the 2 Publication Count metrics, NW043 and NW039. The three rows for each stock show the mean metrics for All data, Training data, and Test data.

Table 2 shows the variability in the sentiment metrics by stock. Some stocks clearly generate a lot of news media attention (such as BHP and ANZ), whilst others generate relatively little (such as CSL, IAG and RIO). This suggests it is appropriate to model each stock using a separate NNAR model. It is also interesting that the company which generates the largest amount of twitter sentiment (TLS) has only a medium amount of news coverage. Therefore, it is

This is a post-peer-review, pre-copyedit version of an article published in Applied Intelligence. The final authenticated version is available online at: http://dx.doi.org/ 10.1007%2Fs10489-019-01458-9.

reasonable to assume that the news sentiment metrics and the twitter sentiment metrics are measuring different things. This is consistent with previous findings in the literature.

Stock Code	Dataset	Observations	mean(News Count)	mean(Twitter Count)
AMC	All	883	17.81	1.4
AMC	Train	863	17.97	1.36
AMC	Test	20	14.95	2
ANZ	All	882	123	25.12
ANZ	Train	862	122.44	23.85
ANZ	Test	20	167.3	65.05
BHP	All	883	430.34	107.63
BHP	Train	863	435.44	110.78
BHP	Test	20	296.6	37.45
BXB	All	883	18.46	0.39
BXB	Train	863	18.67	0.38
BXB	Test	20	14.15	0.55
CBA	All	879	135.88	36.47
CBA	Train	859	135.94	36.16
CBA	Test	20	159.6	59.9
CSL	All	883	14.99	0.85
CSL	Train	863	15.23	0.84
CSL	Test	20	9.8	1.15
IAG	All	883	18.34	2.31
IAG	Train	863	18.4	2.37
IAG	Test	20	19.05	0.55
MQG	All	881	41.77	18.33
MQG	Train	861	41.64	17.99
MQG	Test	20	45.55	26.75
NAB	All	879	57.59	34.57
NAB	Train	859	56.32	34.3
NAB	Test	20	87.3	50.3
ORG	All	878	30.69	3.97
ORG	Train	858	31.09	4.01
ORG	Test	20	19.55	3.7
RIO	All	883	4.04	1.05
RIO	Train	863	4.22	1.09
RIO	Test	20	0	0
S32	All	789	48.62	2.63

S32	Train	769	46.85	2.51
S32	Test	20	88.1	7.95
SUN	All	883	24.9	1.51
SUN	Train	863	25.25	1.51
SUN	Test	20	15.2	1.5
TLS	All	883	47.68	222.5
TLS	Train	863	47.17	221.76
TLS	Test	20	77.2	269.5
WBC	All	880	119.51	43.12
WBC	Train	860	120.4	43.04
WBC	Test	20	105.15	49.1
WES	All	883	34.18	1.63
WES	Train	863	34	1.61
WES	Test	20	40.35	0.65
WOW	All	882	29.01	33.68
WOW	Train	862	29.08	30.76
WOW	Test	20	25.35	107.95
WPL	All	880	79.55	8.9
WPL	Train	860	80.67	9.22
WPL	Test	20	52.4	2

Table 4: Daily Summary Statistics

Two stocks (SCG and TCL) had to be removed due to having no sentiment observations in Bloomberg during the training period, leaving a total of 18 stocks on which to perform modelling. As the sentiment scores are count data, the square root transformation was applied before the scores were submitted as inputs to the NNAR (SENTIMENT) model. Additionally, all inputs to both models followed the standard approach of standardizing by subtracting the column means and dividing by their respective standard deviations.

The NNAR (BASIC) and NNAR (SENTIMENT) models were trained against the in-sample data, and predictions were made on the out-of-sample data. Figures 2 and 3 show a typical example of the quality of the out-of-sample forecasts on WES (Wesfarmers Ltd). Figure 2 shows the in-sample and out-of-sample periods combined to give context, whilst Figure 3 focuses only on the out-of-sample predictions.

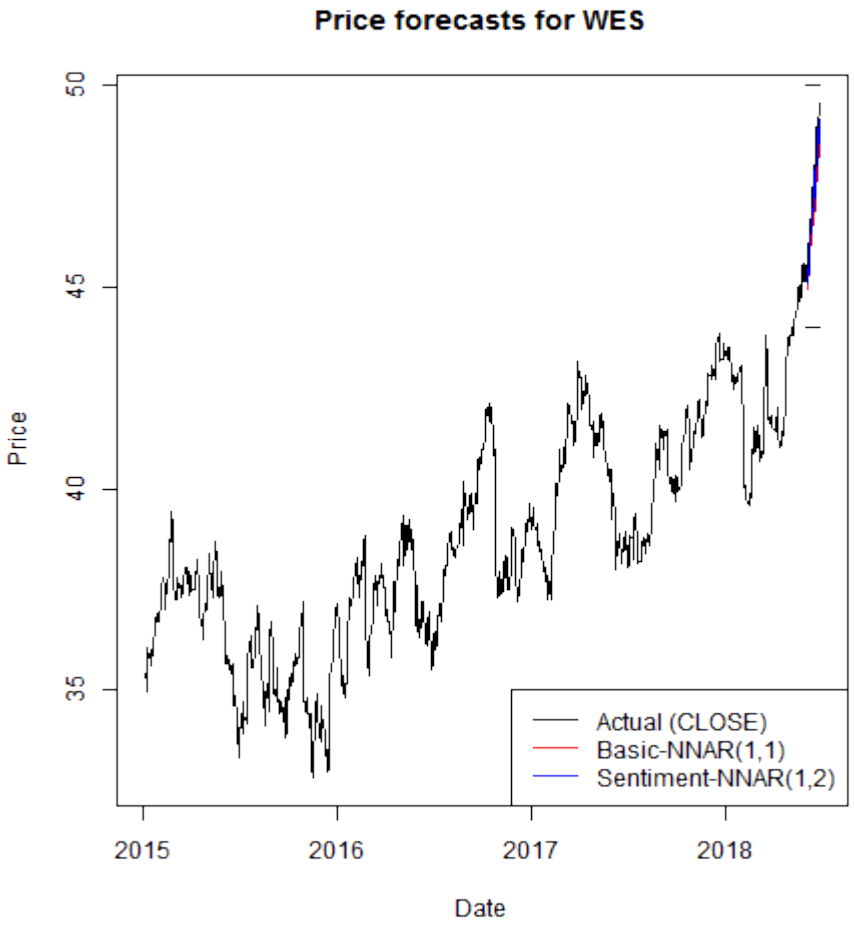

Figure 2: Example of Out-of-Sample prediction (context)

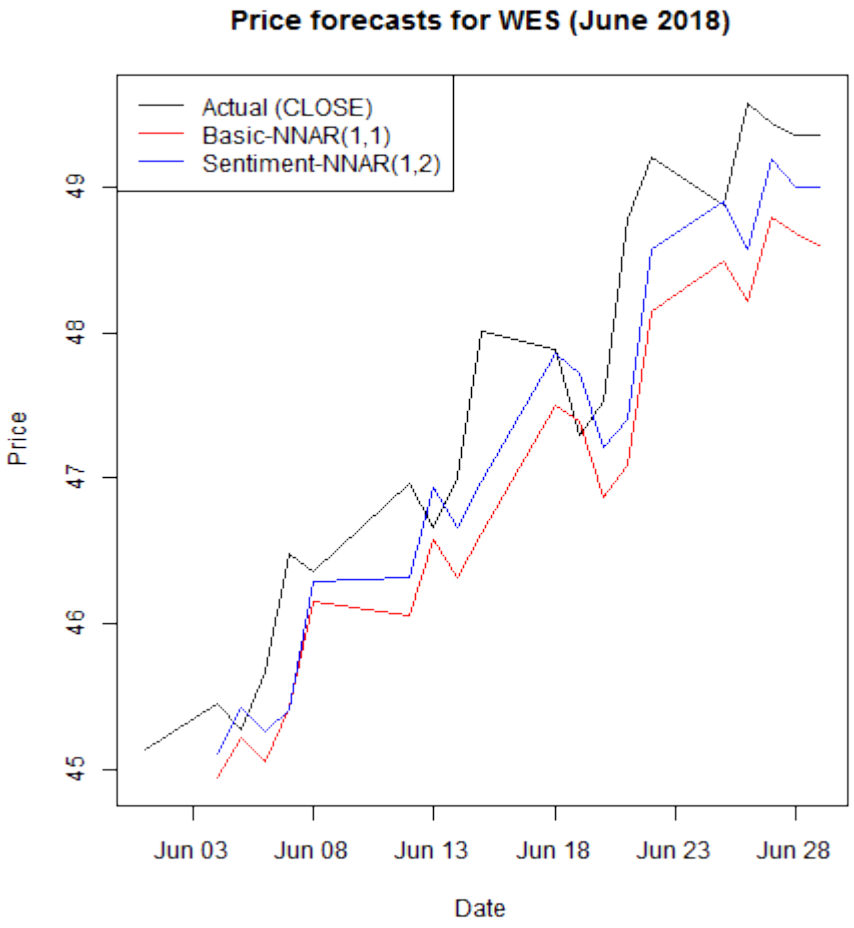

Figure 3: Example of Out-of-Sample prediction

Table 3 shows the overall prediction accuracy results for each stock and model combination. It also shows the RMSE for both the Basic and the Sentiment models on the out-of-sample data, and indicates where the Sentiment model achieved a higher accuracy (as measured by RMSE) than the basic model.

Stock Code	Basic Model	Basic RMSE	Sentiment Model	Sentiment RMSE	RMSE Difference
AMC	NNAR(1,1)	0.1378	NNAR(1,2)	0.1374	0.0004
ANZ	NNAR(2,2)	0.3242	NNAR(2,2)	0.3310	-0.0068
BHP	NNAR(1,1)	0.4359	NNAR(1,2)	0.4206	0.0153
BXB	NNAR(1,1)	0.1213	NNAR(1,2)	0.1252	-0.0039
CBA	NNAR(1,1)	0.9604	NNAR(1,2)	0.9478	0.0126
CSL	NNAR(1,1)	2.4084	NNAR(1,2)	2.3499	0.0585
IAG	NNAR(1,1)	0.1008	NNAR(1,2)	0.0962	0.0045

MQG	NNAR(1,1)	1.9982	NNAR(1,2)	1.8872	0.1110
NAB	NNAR(2,2)	0.2269	NNAR(2,2)	0.2276	-0.0007
ORG	NNAR(1,1)	0.1344	NNAR(1,2)	0.1333	0.0011
RIO	NNAR(1,1)	1.1706	NNAR(1,2)	1.1568	0.0138
S32	NNAR(1,1)	0.0583	NNAR(1,2)	0.0551	0.0033
SUN	NNAR(1,1)	0.1517	NNAR(1,2)	0.1650	-0.0133
TLS	NNAR(1,1)	0.0580	NNAR(1,2)	0.0468	0.0112
WBC	NNAR(2,2)	0.2823	NNAR(2,2)	0.2773	0.0050
WES	NNAR(1,1)	0.8299	NNAR(1,2)	0.6070	0.2229
WOW	NNAR(5,3)	0.4142	NNAR(5,4)	0.3642	0.0500
WPL	NNAR(2,2)	0.5635	NNAR(2,2)	0.5510	0.0126
Mean		0.5765		0.5489	0.0276
Standard Deviation		0.6773		0.6330	0.0554

Table 5: Price Prediction Results

As the forecasts from NNETAR are direct forecasts of price, the same models can be used to test for directional accuracy. Table 6 shows the number of correct directional forecasts (up/down) for each stock made over the out-of-sample period.

Stock Code	Basic	Sentiment	Difference
AMC	8	8	0
ANZ	10	9	-1
BHP	8	7	-1
BXB	6	6	0
CBA	9	9	0
CSL	11	10	-1
IAG	10	12	2
MQG	5	5	0
NAB	7	7	0
ORG	9	8	-1
RIO	8	8	0
S32	10	10	0
SUN	13	13	0
TLS	12	10	-2
WBC	10	10	0
WES	6	8	2
WOW	9	9	0
WPL	7	6	-1

Table 6: Direction Prediction Results

4 Discussion

As Table 3 shows, RMSEs using NNAR models are small, and the difference in RMSE between BASIC and SENTIMENT NNAR models is also small in the majority of cases.

Overall, the sentiment model outperformed the basic model 77.78% of the time in terms of RMSE. There are two appropriate statistical tests to determine whether the SENTIMENT model is statistically significantly better than the BASIC model. As both models make a prediction on every day of the out-of-sample period, the average RMSEs of each model can be compared to each other using a paired samples t-test.

The paired samples t-test null hypothesis is that the true difference in means between the average RMSEs of the SENTIMENT and the BASIC model is zero.

An alternative approach is to use the non-parametric sign test, which treats the data to be tested as a Binomial experiment. In this case, the hypothesis is that the true probability of either outcome is 0.5, or, in other words, either model being the best is equally likely.

A paired-samples t-test was conducted to compare the average RMSEs in the BASIC and SENTIMENT models. There was no significant difference in the scores for the BASIC model and the SENTIMENT model at the 5% level with a test statistic t(17 degrees of freedom) = 2.0557 and a two-sided p-value = 0.055. However, the p-value indicates that there is a difference at any level higher than 5.55%. Furthermore, it could be argued that the normality assumption of the t-test may not be satisfied and so a non-parametric test is more suited. In this case, a Wilcoxon Signed Rank Test is the appropriate non-parametric equivalent and it rejects the null hypothesis that the two models RMSEs are the same with a two-sided p-value of 1.15%. Thus, the RMSE-based results are favourable for the SENTIMENT model.

Results for the signs test also reject the null hypothesis that either model being best is equally likely. The signs test indicates the SENTIMENT model is superior at a 5% level of statistical significance; specifically a signs test with 14 positive results out of a possible 18 yields a two-sided p-value of 0.0308.

It is known that prediction of prices is a difficult problem, and better success has been found predicting stock price direction (the sign of stock returns) by some researchers.

When the models created in this paper are used to predict direction of next price movement (Up/Down), the results are less clear cut. Out of the 18 stocks, the SENTIMENT model direction forecasts are more accurate in 2 stocks, less accurate in 6 stocks, and equal in 10 stocks. The non-parametric Wilcoxon Signed Rank Test is appropriate to use to test the null hypothesis that there is no difference between the number of times that the direction is predicted corrected for the SENTIMENT model as compared with the BASIC model. The test indicates that the null hypothesis is not rejected even at the 10% level and so the results are inconclusive with regard to prediction direction.

From a traders perspective, we propose that this work provides evidence that sentiment has a role to play in stock price modelling. Although more work needs to be done, it is suggested that traders should include total News and Twitter sentiment counts as part of their price discovery process.

This is a post-peer-review, pre-copyedit version of an article published in Applied Intelligence. The final authenticated version is available online at: http://dx.doi.org/ 10.1007%2Fs10489-019-01458-9.

The results of this preliminary study motivate future work that involves modelling and predicting stock returns, direction and volatility using sentiment metrics and a wide variety of machine learning techniques. It is our goal to determine the extent to which sentiment metrics provide additional insight into stock performance.

In an environment where prediction accuracy is of paramount importance, the search for suitable exogenous variables to use as predictors in formal models is a relentless one.

References

Atsalakis, G. S., and K. P. Valavanis, 2009, Surveying stock market forecasting techniques – Part II: Soft computing methods, *Expert Systems with Applications* 36, 5932-5941.

Fama, E., 1965, The Behaviour of Stock Market Prices, *Journal of Business*, 34-105.

Hyndman, R. J., and G. Athanasopoulos, 2013, *Forecasting: principles and practice* (OTexts, Melbourne, Australia).

Long, J. B. D., A. Shleifer, L. H. Summers, and R. J. Waldmann, 1990, Noise Trader Risk in Financial Markets, *Journal of Political Economy* 98, 703-738.

Rapach, D. E., and G. Zhou, 2013, Forecasting stock returns, *Handbook of economic forecasting* 2, 328-383.

Tetlock, P. C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance* LXII, 1139-1168.

Vanstone, B. J., A. Gepp, and G. Harris, 2018, The Effect of Sentiment on Stock Price Prediction, in M. Mouhoub, S. Sadaoui, O. A. Mahamed and M. Ali eds, *Recent Trends and Future Technology in Applied Intelligence* (Springer, Cham) 551-559.

Yen, G., and C.-f. Lee, 2008, Efficient Market Hypothesis (EMH): Past, Present and Future, *Review of Pacific Basin Financial Markets and Policies* 11, 305-329.