Utilizing Sentiment Scores to Forecast Stock Price

Jiaze Wang, Zhilin Song

Thompson Rivers University

DASC 5420

April 14, 2024

Abstract

This study explores the impact of news sentiment on stock price prediction using Natural Language Processing (NLP) techniques. Using a pre-trained RoBERTa model for sentiment scoring for financial news headlines and this is combined with technical factors for stock price prediction. Lasso and Ridge regression models are employed for prediction, with time-series cross-validation used for evaluation. Results show that sentiment scores significantly impact the prediction of certain stocks, such as FXI, but have less influence on others, like Tesla. While the models explain a substantial portion of stock price variation, further research is needed to enhance their robustness.

Keywords: Natural Language Processing, RoBERTa, Sentiment analysis, Stock price prediction, Lasso regression, Ridge regression, Time-series cross-validation

Introduction

Changes in stock prices are of vital importance to those engaged in the financial industry. Therefore, prediction of stock prices has always been a research topic for many practitioners and scholars. However, stock price changes are also affected by many different factors[1], such as war and political factors, etc. Among them, news about the financial field will also have a huge impact on stock price trends, because these news will affect the emotions of investors, and financial decisions are largely driven by emotions[2]. Therefore, we aim to combine the sentiment tendencies in financial news to predict stock prices.

Natural Language Processing (NLP) is a branch of computer science that focuses on the interaction between computers and humans using natural language. It enables computers to understand, interpret, and generate human language, allowing for more natural and intuitive interaction between humans and machines. With advancements in computer technology and the rise of deep learning models, NLP has made significant progress in recent years. It is now widely used in various industries, including finance, healthcare, and many others, to extract valuable insights from large amounts of text data.

Sentiment analysis is a branch of NLP that can extract and analyze emotions and tendencies in text. Emotional tendencies are reflected by emotion scores. In this experiment, sentiment analysis technology will be used as the main text analysis method.

There are also many commonly used models in the field of natural language processing, such as LSTM[3], RNN[4] and Transformer[5]. Among them, the Transformer model is considered to be a deep learning model that can be better applied in the field of NLP. The Transformer model is a deep learning model for processing sequence data, proposed by Vaswani et al. in 2017. When processing sequence data, this model does not need to rely on the order of the sequence like traditional recurrent neural networks (RNN) or long short-term memory networks (LSTM), but uses an attention mechanism to capture the relationship between various positions in the sequence. This also enables the Transformer to better capture long-distance dependencies in sequences.

Because of the good performance of Transformer in NLP tasks, this article selected a pre-training model based on Transformers to conduct sentiment analysis of financial news.

Data

All the links of data that are used in training our natural language model and predicting stock prices are listed in Appendix.

The data set used in this experiment contains financial news from Reuters from January to June 2020 and we selected some stocks such as Tesla and Fxi. These stocks will be used in the task of stock prediction.

An overview of the data can be viewed in the figure 1 below.

number	title	date	stock
530504	US VP Pence Says US Will Continue To Sta	2020-06-10	FXI
1261140	Tesla's Stock Closes At All-Time High As	2020-06-10	TSLA
1261141	'Tesla factory workplace safety is 5% be	2020-06-10	TSLA
1261142	'Tesla hacker unlocks Performance upgrad	2020-06-10	TSLA
1261144	Tesla's Journey To $1,000 In 2020	2020-06-10	TSLA
1261145	Tesla Shares Mark Session And New All-Ti	2020-06-10	TSLA
1261146	Wedbush Says Tesla Has 'Game Changing' D	2020-06-10	TSLA
1261147	Wedbush Maintains Neutral on Tesla, Rais	2020-06-10	TSLA
1261148	Musk Says It's Time To Ramp Production C	2020-06-10	TSLA
1261149	Tesla shares are trading higher after We	2020-06-10	TSLA
1261150	Electrek Reports Tesla VP Of Business De	2020-06-10	TSLA
1261151	Wedbush Raises Tesla Price Target From $	2020-06-10	TSLA
1261152	Tesla's Musk Says It's Time To Bring Tes	2020-06-10	TSLA
1261153	Hearing Wedbush Raises Tesla Base Price	2020-06-10	TSLA
1261154	Key Tesla Executive Behind Shanghai Giga	2020-06-10	TSLA
530508	China May CPI -0.8% MoM vs -0.5% Est	2020-06-09	FXI
530509	'The State Department has informed Congr	2020-06-09	FXI
1261155	Volkswagen Cuts Costs, Targets Electric	2020-06-09	TSLA
1261156	Tesla Bolsters EV Enthusiasm With German	2020-06-09	TSLA
1261157	Tesla Bull Vs. Bear: Ron Baron And Paul	2020-06-09	TSLA
530510	Morgan Stanley CEO James Gorman Says Gre	2020-06-09	FXI
1261158	'Tesla adds Model Y to referral program,	2020-06-09	TSLA
530512	Hearing Investor Kyle Bass Considering L	2020-06-09	FXI
530514	WHO's Ryan Says 'Absolutely Convinced' T	2020-06-09	FXI
530515	WHO's Van Kerkhove Says Some Modelers Es	2020-06-09	FXI
1261161	UPDATE: 'Tesla defied county orders so i	2020-06-09	TSLA
1261162	Tesla Reported Several Coronavirus Cases	2020-06-09	TSLA

Figure 1. The overview of the Reuters news data

In addition, in order to enable the model to better understand financial vocabulary and determine the emotional tendency of financial news, the Loughran-McDonald financial dictionary is also used in conjunction with the FIN111 data set to train the model.

The financial dictionary contains information including the number of dictionary reports, the proportion of the total, the average proportion of each document, the standard deviation of the proportion of each document, the number of documents, and 7 sentiment category identifiers. The FIN111 data set is a data set that contains numerous financial news articles collected from well-known financial publications. These data are divided into three categories according to their impact on the financial market, namely weak, semi-strong and strong. We selected 6,000 pieces of data to train the model in combination with the financial dictionary to improve the performance of the model.

An overview of the dictionary data can be viewed in the figure 2 below.

Word	Seq_num	Word Cou	Word Pro	Average P	Std Dev	Doc Coun	Negative	Positive	Uncertain	Litigious	Strong_M	Weak_Mo	Constraini	Syllables	Source
AARDVAR	1	354	1.55E-08	1.42E-08	3.82E-06	99	0	0	0	0	0	0	0	2	12of12inf
AARDVAR	2	3	1.31E-10	8.65E-12	9.24E-09	1	0	0	0	0	0	0	0	2	12of12inf
ABACI	3	9	3.94E-10	1.17E-10	5.29E-08	7	0	0	0	0	0	0	0	3	12of12inf
ABACK	4	29	1.27E-09	6.65E-10	1.60E-07	28	0	0	0	0	0	0	0	2	12of12inf
ABACUS	5	8570	3.75E-07	3.81E-07	3.53E-05	1108	0	0	0	0	0	0	0	3	12of12inf
ABACUSES	6	0	0	0	0	0	0	0	0	0	0	0	0	4	12of12inf
ABAFT	7	4	1.75E-10	2.30E-11	2.46E-08	1	0	0	0	0	0	0	0	2	12of12inf
ABALONE	8	142	6.22E-09	4.97E-09	1.07E-06	48	0	0	0	0	0	0	0	4	12of12inf
ABALONE	9	1	4.38E-11	8.28E-11	8.85E-08	1	0	0	0	0	0	0	0	4	12of12inf
ABANDON	10	127090	5.56E-06	4.70E-06	3.31E-05	66312	2009	0	0	0	0	0	0	3	12of12inf
ABANDON	11	234345	1.03E-05	1.08E-05	7.68E-05	107787	2009	0	0	0	0	0	0	3	12of12inf
ABANDON	12	20962	9.18E-07	8.42E-07	1.20E-05	13112	2009	0	0	0	0	0	0	4	12of12inf
ABANDON	13	285978	1.25E-05	1.24E-05	8.16E-05	94669	2009	0	0	0	0	0	0	4	12of12inf
ABANDON	14	15433	6.76E-07	9.31E-07	2.26E-05	6517	2009	0	0	0	0	0	0	4	12of12inf
ABANDON	15	7616	3.33E-07	1.70E-07	3.65E-06	5457	2009	0	0	0	0	0	0	3	12of12inf

Figure 2. An overview of the Loughran-McDonald financial dictionary

An overview of the FIN111 dataset is shown below.

ID	Headline	Descriptio	Publicatio	Source	Semi-Stro	Strong	Weak
385	firms_ unwise a	a bumper	24/08/202	Economist	1	0	0
4	why young con	the ancier	12/10/202	Economist	1	0	0
9	why companies	the unlike	05/10/202	Economist	1	0	0
13	bill ackman wa	a new spir	05/10/202	Economist	1	0	0
14	so long iphone.	is this the	05/10/202	Economist	0	0	0
15	america__ bosse	will lawsui	02/10/202	Economist	1	0	0
21	hollywood__ str	next: the a	25/09/202	Economist	1	0	0
28	what arm and i	the old-sc	21/09/202	Economist	1	0	0
30	could openai b	what the I	18/09/202	Economist	1	0	0
31	arm__ successfu	but risks s	14/09/202	Economist	1	0	0
34	electric two-wh	cross-bor	14/09/202	Economist	1	0	0
36	chinese carmak	the eu lau	14/09/202	Economist	0	0	1

Figure 3. An overview of the FIN111 dataset

Data preprocessing

Preprocessing is to enable the original text to be better processed and transformed into an input form that the model can process. These include removing special characters, segmenting words, marking the beginning and end of text, converting to numbers, padding or truncation to ensure that the text being processed is of the same length, and generating masks to mark real text and padded text.

For the data in Loughran-McDonald financial dictionary, we classify emotions into three categories: positive, neutral, and negative, represented by the numbers 2, 1, and 0 respectively.

For the FIN111 data set, punctuation removal, stop word removal, spelling proofreading, word segmentation and phrase detection were performed.

Stock data are aligned with news events and the corresponding sentiment score based on trade dates. Any null values are dropped. The three categories of sentiment score are further transformed into one value by the formula (2) below.

Method

Sentiment analysis

In the task of sentiment analysis, the pre-trained RoBERTa[6] was selected as the basic model. RoBERTa is a pre-trained language model based on Transformer proposed by Facebook AI. RoBERTa was pre-trained on a large-scale text corpus and learned a universal language representation. RoBERTa uses a bidirectional Transformer encoder that can simultaneously consider the contextual information of each position in the text, which allows it to better understand the meaning of the text. So its performance on NLP tasks is better than other traditional models.

The model used in this experiment is a pre-trained model, which is a model trained for sentiment analysis using Twitter data. The model then combines the Loughran-McDonald financial dictionary with data from the FIN111 dataset to learn sentiment analysis and scoring of news headlines.

We matched the word-segmented FIN111 data set with the financial dictionary, and assigned labels to words that appeared in both data sets to determine the emotional tendencies of the words, and used the obtained data set to train the model.

The parameter settings during model training are as follows:

The training batch size is 16, the number of training times is 3, and the proportion of the warm-up phase to the total training steps is set to 10%, which means that the first 10% of the steps are used for warm-up, and the type of learning rate scheduler is linear scheduler.

Finally, the model combined with financial dictionary learning was used to score Reuters news related to the selected stocks.

The overview of the News results after scoring is shown below.

number	title	date	stock	sentiment_scores
530504	US VP Pence Says US Will Continue To Sta	2020-06-10	FXI	{'negative': 0.0027036332, 'neutral': 0.99255544, 'positive': 0.004740865}
1261140	Tesla's Stock Closes At All-Time High As	2020-06-10	TSLA	{'negative': 0.61769533, 'neutral': 0.37977692, 'positive': 0.0025278318}
1261141	'Tesla factory workplace safety is 5% be	2020-06-10	TSLA	{'negative': 0.0029329613, 'neutral': 0.93131495, 'positive': 0.06575208}
1261142	'Tesla hacker unlocks Performance upgrad	2020-06-10	TSLA	{'negative': 0.0022431153, 'neutral': 0.9911049, 'positive': 0.0066520753}
1261144	Tesla's Journey To $1,000 In 2020	2020-06-10	TSLA	{'negative': 0.0013240529, 'neutral': 0.98983824, 'positive': 0.008837665}
1261145	Tesla Shares Mark Session And New All-Ti	2020-06-10	TSLA	{'negative': 0.0014801671, 'neutral': 0.9945475, 'positive': 0.003972349}
1261146	Wedbush Says Tesla Has 'Game Changing' D	2020-06-10	TSLA	{'negative': 0.0025372084, 'neutral': 0.9949202, 'positive': 0.0025426717}
1261147	Wedbush Maintains Neutral on Tesla, Rais	2020-06-10	TSLA	{'negative': 0.0033200346, 'neutral': 0.99488974, 'positive': 0.001790174}
1261148	Musk Says It's Time To Ramp Production C	2020-06-10	TSLA	{'negative': 0.0018354147, 'neutral': 0.9966543, 'positive': 0.0015103269}
1261149	Tesla shares are trading higher after We	2020-06-10	TSLA	{'negative': 0.0015484956, 'neutral': 0.9910046, 'positive': 0.0074469335}
1261150	Electrek Reports Tesla VP Of Business De	2020-06-10	TSLA	{'negative': 0.008167173, 'neutral': 0.991074, 'positive': 0.00075886096}
1261151	Wedbush Raises Tesla Price Target From $	2020-06-10	TSLA	{'negative': 0.0021347785, 'neutral': 0.9943051, 'positive': 0.0035601691}
1261152	Tesla's Musk Says It's Time To Bring Tes	2020-06-10	TSLA	{'negative': 0.0019563367, 'neutral': 0.99627566, 'positive': 0.0017680377}
1261153	Hearing Wedbush Raises Tesla Base Price	2020-06-10	TSLA	{'negative': 0.010186483, 'neutral': 0.98914367, 'positive': 0.0006698685}
1261154	Key Tesla Executive Behind Shanghai Giga	2020-06-10	TSLA	{'negative': 0.052904766, 'neutral': 0.945385, 'positive': 0.0017101511}
530508	China May CPI -0.8% MoM vs -0.5% Est	2020-06-09	FXI	{'negative': 0.0055675097, 'neutral': 0.9933882, 'positive': 0.0010442795}
530509	'The State Department has informed Congr	2020-06-09	FXI	{'negative': 0.0022227622, 'neutral': 0.99644333, 'positive': 0.0013339284}
1261155	Volkswagen Cuts Costs, Targets Electric	2020-06-09	TSLA	{'negative': 0.0033000421, 'neutral': 0.99420506, 'positive': 0.0024949622}
1261156	Tesla Bolsters EV Enthusiasm With German	2020-06-09	TSLA	{'negative': 0.0022379924, 'neutral': 0.9935309, 'positive': 0.0042310474}
1261157	Tesla Bull Vs. Bear: Ron Baron And Paul	2020-06-09	TSLA	{'negative': 0.0025280397, 'neutral': 0.99667656, 'positive': 0.0007954875}
530510	Morgan Stanley CEO James Gorman Says Gre	2020-06-09	FXI	{'negative': 0.0027301805, 'neutral': 0.9960842, 'positive': 0.0011855953}
1261158	'Tesla adds Model Y to referral program,	2020-06-09	TSLA	{'negative': 0.18254709, 'neutral': 0.81477433, 'positive': 0.0026785433}
530512	Hearing Investor Kyle Bass Considering L	2020-06-09	FXI	{'negative': 0.36176258, 'neutral': 0.63684887, 'positive': 0.0013885947}
530514	WHO's Ryan Says 'Absolutely Convinced' T	2020-06-09	FXI	{'negative': 0.021978509, 'neutral': 0.9768844, 'positive': 0.0011370655}
530515	WHO's Van Kerkhove Says Some Modelers Es	2020-06-09	FXI	{'negative': 0.44396546, 'neutral': 0.5543631, 'positive': 0.0016714986}
1261161	UPDATE: 'Tesla defied county orders so i	2020-06-09	TSLA	{'negative': 0.82943803, 'neutral': 0.1685476, 'positive': 0.0020143543}

Figure 4. The overview of the scored Reuters news data

Stock Price Prediction with Sentiment Score

Based on the sentiment scores above for the news of the two selected stocks (Tesla and Fxi), the closing prices of the two stocks can be predicted. There are three categories of sentiment scores, the original formula[7] to combine all three categories and yield one score,

$$senfiment\_score = \frac{postive - negative}{postive + negative} \quad (1)$$

We further improve this sentiment formula by incorporating neutral score into it, which is

$$senfiment\_score = \frac{postive - negative}{postive + negative}(1 - neutral) \quad (2)$$

The neutral score shrinks the formula (2), this matches our intuition that neutral news tends to minimize stock price movement. The higher the positive score, the lower the negative score will lead to a higher the sentiment-score, and hence the more possibility of positive price movement will happen, vice versa. Since the three categories of scores sum to one, and each of them are positive, hence the value range for sentiment score is between -1 and 1.

In addition to sentiment scores, some other technical factors such as Simple Moving Average, Relative Strength Index, Moving Average Convergence Divergence, Bolling Band are calculated by python package Talib, they are also included in our model to improve the predictability. While there are more than 5 factors in our model, we try to avoid the overfitting problem by using Lasso and Ridge regression. And since stock data is time-series data, the time-series specific cross-validation techniques are employed to test the model performance. In this project, we use 5-folds time series cross-validation. The lasso and ridge regression hyperparameter alphas are tuned between 0 and 100.

In order to test the effectiveness of sentiment scores, we compare the performance of the original lasso and ridge model with another model that contains all the same factors as the first model except sentiment scores.

Github Link: https://github.com/wangjiaze/Utilizing-Sentiment-Scores-to-Forecast-Stock-Price

Results

For the stock FXI, as Figure 5 presents, most of the sentiment scores are around zero, there is no clear pattern from the graph.

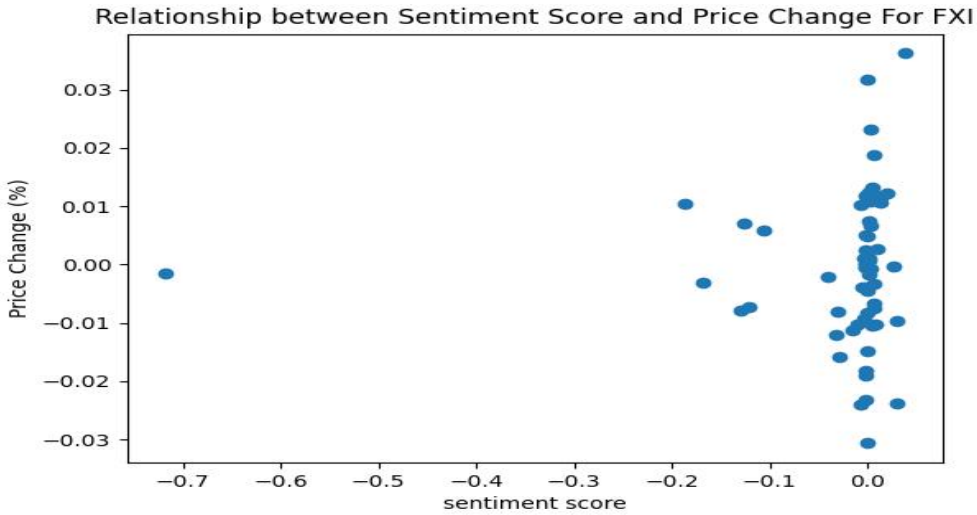

Figure 5. Relationship Between Sentiment Score and Price Change For Fxi

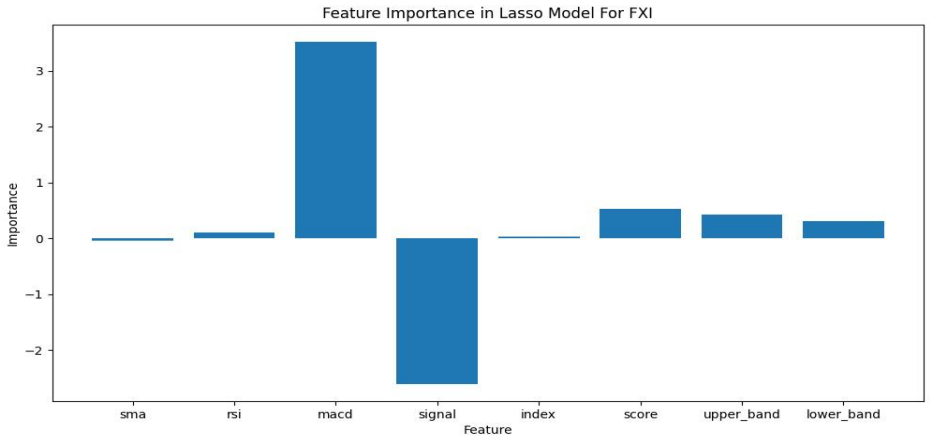

Figure 6. Feature Importance In Lasso Model For Fxi

As the above graph presents, sentiment score is the third most important factor in the Lasso model with a corresponding value of 0.531. This means that for every one unit increase in score, the estimated stock closing price for Fxi will increase 0.531. If considering all the given factors, the Lasso model can explain about 57 percent of the variance in the closing price of Fxi, but the R-square of Lasso decreases about 5 percent without the sentiment score. However, the feature importance of sentiment score is only about 0.02 percent when applying Ridge regression. But there is about 6 percent drop in the R-squares for Ridge regression when we exclude the sentiment score factor. Ridge regression outperforms Lasso with only 0.2 higher R-squares. The R-squared results are shown

in Table 1 below. and both models can capture most of the price movement trend as shown in Figure 7, and Figure 8.

Table 1. R-squares For Lasso And Ridge Regression With And Without Sentiment Scores

	Lasso With Sentiment Score	Lasso Without Sentiment Score	Ridge With Sentiment Score	Ridge Without Sentiment Score
Fxi	0.5695	0.5236	0.5716	0.5137
Tesla	0.4619	0.5988	0.4724	0.6086

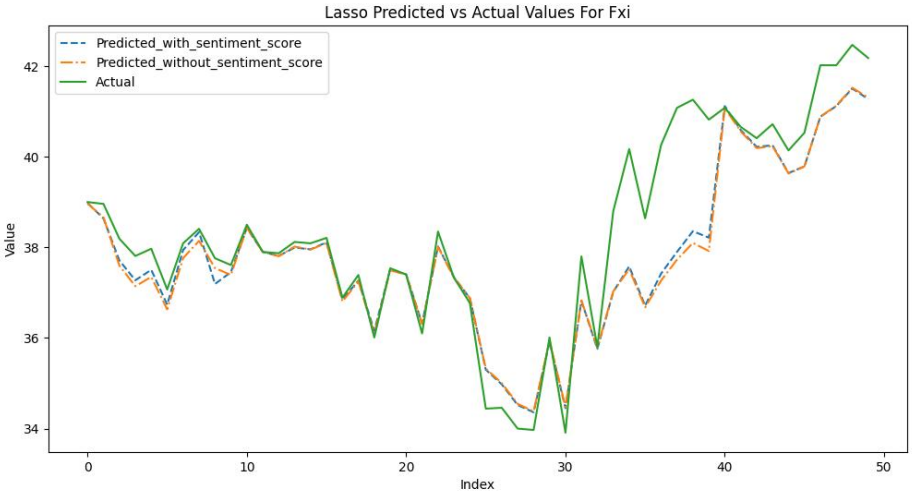

Figure 7. Actual and Lasoo Predicted Closing Price For Fxi

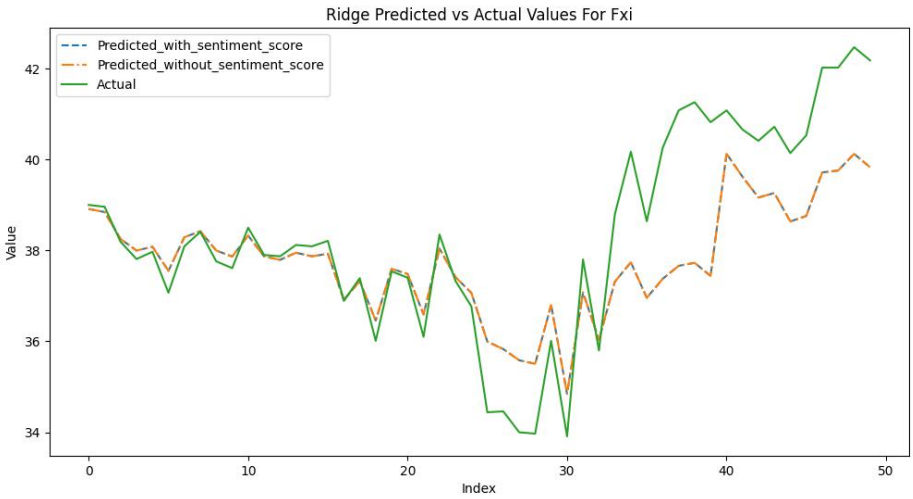

Figure 8. Actual and Ridge Predicted Closing Price For Fxi

For the stock Tesla, as Figure 9 presents, most of the sentiment scores are around zero, there is no clear pattern from the graph, but the sentiment score for Tesla is more spread.

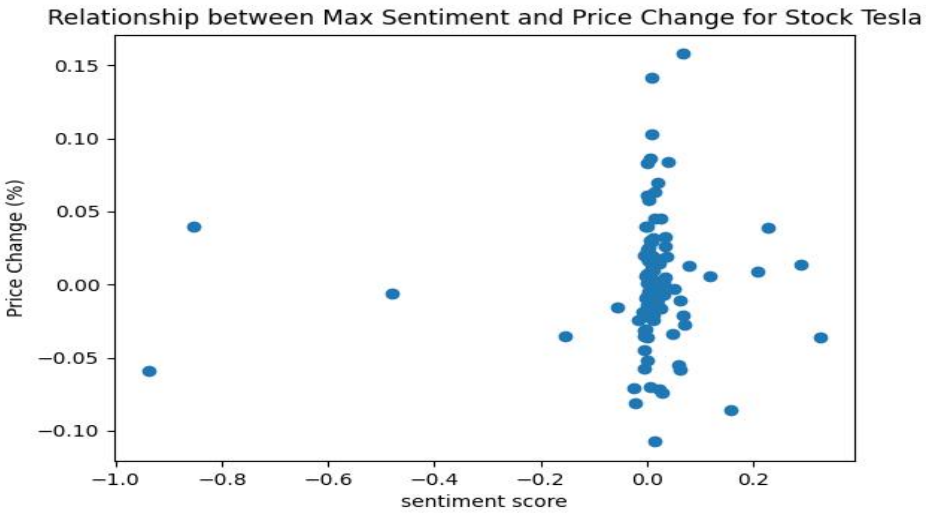

Figure 9. Relationship Between Sentiment Score and Price Change For Tesla

Figure 10. Feature Importance In Lasso Model For Tesla

As the above graph presents, sentiment score is the fifth most important factor in the Lasso model with a corresponding value of 0.17, which means the estimated stock closing price for Tesla will increase 0.17 for every one unit increase in sentiment score.

By including sentiment score in Lasso, the model's R-square decreased from 60 percent to 46 percent, this time it is not necessary to include the sentiment score. For the Ridge regression model, it is the same situation that R-square decreases from 60 percent to 47 percent after including the sentiment score. Therefore, the sentiment score factor performs poorly in Tesla close price prediction, but other factors can capture the main price movement tendency as shown in Figure 11 and 12.

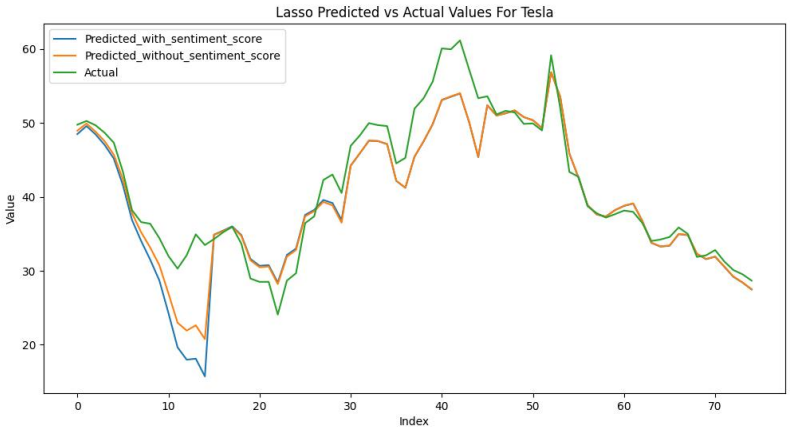

Figure 11. Actual and Lasoo Predicted Closing Price For Tesla

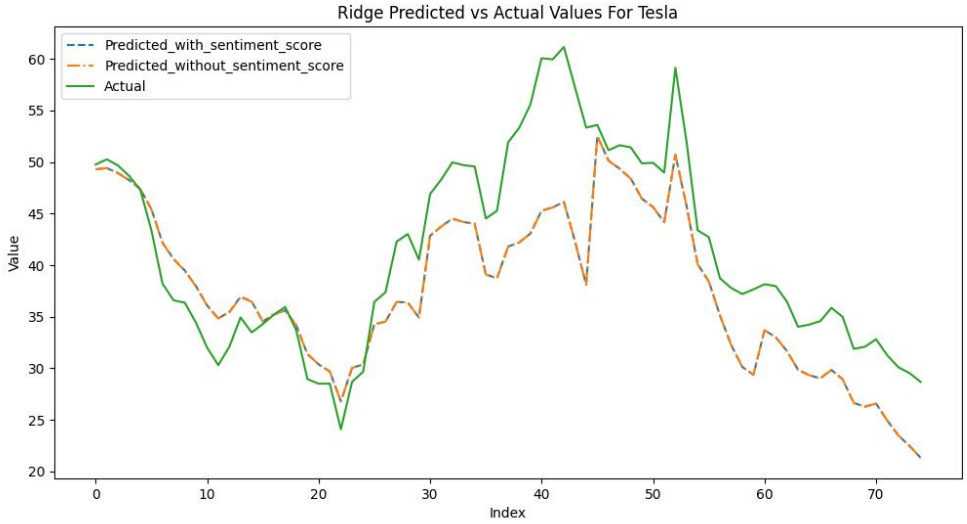

Figure 12. Actual and Ridge Predicted Closing Price For Tesla

Discussion

The sentiment score factor constructed by our natural language processing model is the third significant factor to predict Fxi closing price in the Lasso model with a coefficient value of 0.531, and the R-square of Lasso increases 5 percent by including sentiment score. This factor also helps increase about 6 percent the R-square of ridge regression. However, sentiment factors fail in predicting Tesla close price in both of the Lasso and Ridge regression, which result in more than 10 percent decrease in the corresponding R-squares. Figure 11,12 also demonstrates that there is nearly no change on the regression line by including the sentiment factor.

All the other factors other than sentiment score can capture most of the price movement trend for both stocks and explain about 60 percent of variation in Tesla closing price, 57 percent of variation in Fxi closing price.

This result is not robust because of the limited amount of stock data and news used. Further improvement can be done through using different formulas to construct the sentiment score, including more testing stocks and more amount of trading data for each stock.

References

[1]. Agrawal J, Chourasia V, Mittra A. State-of-the-art in stock prediction techniques. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. 2013;2(4):1360–1366.

[2]. Nofsinger JR. Social Mood and Financial Economics. The Journal of Behavioral Finance. 2005;6(3):144–160.

[3]. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735–1780.

[4]. Williams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E. Learning representations by back-propagating errors. Nature. 1986, 323 (6088): 533–536.

[5]. Vaswani A, Shazeer N, Parmar Niki , Uszkoreit Jakob , Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arxiv. 2017 [cited 2024 Apr]. Available from: https://arxiv.org/abs/1706.03762.

[6]. Liu Y, Ott M, Goyal Naman , Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet]. arxiv. 2019 [cited 2024 Apr]. Available from: https://arxiv.org/abs/1907.11692?context=cs.CL.

[7]. Nguyen TH, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications. 2015;42(24):9603-9611.

Appendix

Loughran-McDonald_MasterDictionary:
https://sraf.nd.edu/loughranmcdonald-master-dictionary

FIN 111:
https://github.com/WilliamBeckhauser/FIN111K/blob/main/FIN111K.csv

Reuters Financial News used for scoring:
https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests/data

Tesla Stock Data:
https://ca.finance.yahoo.com/quote/TSLA/history?period1=1577923200&period2=1591833600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

Fxi Stock Data:
https://ca.finance.yahoo.com/quote/FXI/history?period1=1581984000&period2=1591833600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true