PittCoin Final Report

Kartik, Brian, Kasvitha, Owen
kah481, btl26, kak600, omw3
University of Pittsburgh

1 Introduction

Predicting stock movement and future stock prices is a very difficult task, as it is influenced by many factors including the economy, public opinion, company restructuring, and many more; however, accurately predicting future prices could lead to high profits. For this reason, there is a high demand for such technology across businesses and investors.

The goal of this project is to contribute a novel approach for predicting stock prices using financial and business related news articles more accurately, specifically focusing on the S&P500 Index. This project aims to answer two main questions:

1. Can we predict the credibility of financial news articles?

2. Can we more accurately predict stock price movement if we have an article and its credibility rather than just the article?

We answer the above questions by providing an exploratory analysis of the credibility of news sources vs. the resulting stock movement and use these results to predict stock price movement. We are proposing a new approach to stock prediction models with credibility as an additional input feature, something that previous approaches don't consider.

2 Previous Work

There have been many different approaches for predicting stock movements, with most approaches being defined by hardware limitations, data limitations, or the modern machine learning approaches during their time. Some early approaches used a naive Bayesian classifier to classify if the stock will move up, down, or not at all [1], while more modern approaches incorporate neural networks, specifically RNNs and LSTMs, to better capture movement in time-series data [2 & 3]. While many of these papers have trained their models using exclusively large, credible news outlets [1, 2, 3], other papers have used posts scraped from Twitter and other social media platforms [4, 5]; however, none of these papers use this difference in source credibility as a feature to predict the stock movement.

It is important to note that all of the approaches listed above use S&P500 stock prices for the data input and more recent approaches average to a context window of ten days. Given that these approaches do not assign articles to a specific stock, but rather aggregate the articles to predict all S&P500 stocks individually, we are able to generalize the same techniques to the S&P500 index fund.

There has been work done specifically looking at fake news and stock prices to analyze market manipulation [6], but this fails to analyze the difference between fake and verified news and their differing impact on the market. Additionally, this work is conducted with a private dataset of verified fake news through an SEC investigation, so there is a lack of quality, publicly-available data for credibility of news articles versus stock price movement. Another study was conducted specifically on predicting whether financially relevant tweets were credible using a variety of models, with RF and kNN models being the most successful [7]; however, this study also failed to take the next step and apply the predicted credibilities to market predictions.

After looking through this related work, there is a unique and interesting opportunity to analyze the intersection of these two problems and understand how the credibility of news differently impacts stock price movement, and thus use this information and credibility classification to better predict stock price movement and improve these models.

3 Data

Because the project is separated into two main tasks, credibility classification and stock price prediction, we have used different datasets and preprocessing techniques for each of these tasks. We will breakdown the data we've used below.

3.1 Stock Prediction Data

The data for our stock prediction models were taken from Kaggle [11] and consists of US financial news articles from January to May 2018. These articles were sourced from Bloomberg.com, CNBC.com, reuters.com, wsj.com, and fortune.com. We then took the Title, Text, Date, and Publisher of each article in the dataset and assigned the associated stock prices. We used the YahooFinance API to obtain the S&P 500 Index closing price for a given day. Each article was then assigned the stock prices within a ten day window, including just the closing price of ten days before the article's date to ten days after the article's date. We then calculated a 'Change' field to signify if the index went up or down after the article was published. This change field was calculated by taking the difference between the most recent closing price leading up to the article's published date and the most recent closing price after the published date. This approach allows us to use articles that were published on weekends or holidays where they don't have a closing price.

3.2 Credibility Data

In order to maximize the performance of the model and thoroughly test the benefit of credibility, we explored multiple ways of assigning credibility to the articles. These different approaches required different datasets. For the model-based approach we used two datasets from kaggle [8 & 9] and initially did some simple preprocessing to make a large table with title, text, and credibility label columns. After the first model performed extremely well on the test credibility data but poorly on the stock dataset (very low average credibility score despite articles being at least real, if perhaps not highly credible), we decided to filter our data in this training stage of the first model to be more similar to the language of the data used for the second model. This involved keeping only rows of the dataset that had certain financial terms in their dataset in an attempt to limit the training space to financial news. For the domain-based approach, we used the domain quality ratings by Lin [10] and used the publisher's domain to assign credibility.

4 Methodology

The implementation of our system consists of two models, one of which is a credibility classifier that assigns a credibility score to a given article. The other model is a stock prediction model that will predict if the S&P500 index will increase or decrease.

Both of these models were made by fine-tuning distilbert-base-uncased from HuggingFace as well as the standard tokenizer for this model. Distilbert was chosen as it is a high performing transformer for classification tasks but faster and more lightweight than its counterpart, BERT.

For the credibility model, 'title' and 'text' fields were concatenated with a separator token and the model was trained to predict the 'label' column. The full hyperparameter specifications can be found in the credibility training python notebook in the github repository for this project.

The baseline stock-prediction model took a single string input that consisted of the *title*, *text*, and previous ten day stock prices, all concatenated together with a separator token. This concatenated string was the single input to then predict the *change* column. With our new approach, we have the same input sequence, but also concatenate the credibility score to the front of the input, also with a separator token. By keeping the models very similar with only one minor change, it limits the variables in the experiment and allows us to test the effect of the confidence specifically.

5 Results

5.1 Credibility Model Results

The credibility classification model was very successful on the credibility test data. Initially after training, the model had the results seen in the first row of Table 1, which an accuracy over 99%; however, when the model was used to predict the credibility class probability of each of the financial articles to be used for the stock prediction task, the average probability was extremely low.

In an attempt to improve the model output usability in the stock prediction task, the credibility training dataset was trimmed of any entries that did not contain any financial terminology (terms like stock, ETF, VWAP), so that the model would train on data more similar to the data it would be

	Accuracy	Precision	Recall	F1
Credibility Model v1	0.9929	0.9929	0.9929	0.9929
Credibility Model v2	0.9195	0.9195	0.9195	0.9195

Table 1: Performance metrics of credibility classification model.

Title	Text	Credibility
U.S. judge orders government to release Iraqis or grant bond hearings	A U.S. judge ordered the government on Tuesday to either release Iraqi immigrants it arrested last year or grant them bond hearings, in the latest judicial curb on the Trump administration's efforts to tighten U.S. immigration. Last year the federal government detained hundreds of Iraqi immigrants who had been ordered deported years ago due to criminal convictions. Iraq until recently had refused to take them back, but struck a deal with the United States in March to repatriate its citizens, sparking the immigration sweeps.	0.07329
Cramer reflects on how Trump's actions are fueling the 'beast' market rally.	Cramer reflects on how Trump's actions are fueling the 'beast' market rally 1 Hour Ago Jim Cramer examined the notion that investors are "bored" with the market rally and explained how the president is driving stocks higher.	0.005587
The Wall Street Journal: Peter Thiel's VC firm has made a monster bet on bitcoin	One of the biggest names in Silicon Valley is placing a moonshot bet on bitcoin BTCUSD, +0.72% .Founders Fund, the venture-capital firm co-founded by Peter Thiel, has amassed hundreds of millions of dollars of the volatile cryptocurrency, people familiar with the matter said. The bet has been spread across several of the firm's most recent funds, the people said, including one that began investing in mid-2017 and made bitcoin one of its first investments. Founders and Thiel, 50 years old, are well-known for early investments in companies like Facebook Inc. FB, +2.81% that sometimes take years to come to fruition. The bitcoin bet is quickly showing promise. Founders bought around $15 million to $20 million in bitcoin, and it has told investors the firm's haul is now worth hundreds of millions of dollars after the digital currency's ripping rise in the past year.	0.003425
Emerging markets are set for an even bigger rally in 2018, says one technician	Emerging markets soared more than 33 percent in 2017, and Todd Gordon of TradingAnalysis.com says the rally won't stop. A big part of the rally in emerging markets, tracked by the emerging market ETF EEM , was a weak dollar. And given that Gordon still sees the inverse relationship between EEM and the dollar, measured in his charts by the dollar-tracking ETF UUP , he believes the U.S. currency will continue to help the group. "We have a falling U.S. dollar, which will support international and emerging market currencies and will give those EEM stocks a boost," Gordon said Tuesday on CNBC's "Trading Nation." The U.S. dollar in 2017 posted its worst annual performance in 14 years, while EEM saw its best performance since 2013. As for how high the latter could go, Gordon says EEM has broken "resistance" at around $45, which was the ETF's 2014 highs. That $45 region is now what he calls "support," and he sees it rallying to $50, which the ETF hasn't hit since mid-2011. To play for a move higher, Gordon suggested buying the February 48/50 call spread for 72 cents, or $72 per options contract. This means that if EEM closes above $50 on Feb. 16, then Gordon could make a maximum reward of $128 on the trade. But if EEM were to close below $48, then Gordon would lose the $72 he paid for the trade. As a result, Gordon wants to establish a point at which to get out. "If the 72 cent premium we just laid out gets cut in half to about 36 cents, let's cut the trade and move on," he said. EEM started the year off strong, rallying more than 1 percent on Tuesday.	0.004278

Table 2: Credibility predictions on random article examples from stock training set.

used on later. This model's performance on the credibility test set can be seen in the second row of Table 2. This model had lower accuracy, but at almost 92% did perform quite well. The credibility prediction probabilities were much higher in this case and had a higher spread, but the average predicted probability of an article being credible was still only 0.07567. A sample of the credibility predictions on a few of our articles can be seen in table 2.

Using table two, we can start to speculate on some of the issues with this model and its use with the stock prediction model. These 4 rows of the dataset were randomly selected, and the first 2 mention a prominent political figure (Trump), and 2-4 are all very bullish and energetic analyses of a particular strategy in a certain investment domain. A large portion of fake and low credibility news focuses on political topics, and this section of news was particularly large in the credibility training set, and even when the training set was limited to financial news, fake news examples were often still politically focused. It is likely that the model has learned to classify a disproportionate amount of articles that mention these political figures that were so prominent in fake datasets as fake. Also common in the articles labeled as fake in the training set are articles that have too much of one sentiment, use too many or too few emotional words, or make grandiose claims; however, the nature of a lot of these financial opinion pieces that make up so much of our stock dataset is that people are writing articles about specific things they believe will make them a lot of money in the near future, so they are very energetic when discussing them. These two trends could be responsible for the overwhelming majority of low credibility labels assigned to the input texts for our stock model.

5.2 Stock Prediction Results

As we can see in Table 3, the baseline model had the best performance, while the models that included the credibility of the articles performed slightly lower. After trying two different approaches for assigning credibility, and after multiple iterations of the credibility classification model, the performance is consistently lower than the baseline. This drop in performance may be caused by large buzz words or controversial topics being present in the article. In Table 4, we can see a random sample of articles that were predicted correctly by the baseline, but predicted incorrectly when we add credibility. In many of these examples we see words like 'China' and 'Trump,' which are very controversial topics that are more likely to be used within fake news and misinformation.

We also created a model that is only trained on the stock prices and a model that's only trained on the articles in order to see what the most important feature was in the baseline model. As we can see, the article-only model significantly outperformed the stock-only model, which leads us to conclude that the baseline model almost exclusively uses the articles to make the predictions (Table 3). This may reflect some limitations in the training data, which will be discussed in the next section.

6 Challenges and Limitations

Much of the challenges with our experiments come from a lack of data, which we believe contributes to the results and performance of our models. For news articles and stock prices, our dataset consists of over 300,000 articles, but only span about 90 days. Many of the baseline approaches and previous work setup their experiments where each row is a date, and all of the articles are aggregated on that date. In our setup, we only span 90 days, so aggregating by date would give too few data samples to accurately train a model. Additionally, the volume of articles would contain too many tokens for our model inputs and available computing resources. This evaluation is why we setup the experiment with each row as it's own article; however this may explain why the baseline is heavily relying on the article. We have found that many of the articles include the date within it, meaning that these dates are repeated over different data samples and may be inflating the performance. We intend to remove the dates from the articles as an additional preprocessing step in order to resolve this issue in the future.

The limited data available likely also impacted the model results when including the credibility score of the articles. For the domain-based credibility score, we assigned credibility by domain, however our dataset only consisted of articles from five domains, two of which shared the same credibility score. This limited variation in the credibility between domains may have decreased the model performance. Additionally, our model-based credibility classification was trained on a different dataset than our stock prediction model. This disconnect may have led to many incorrect credibility

	Accuracy	Precision	Recall	F1
Baseline Model	0.8052	0.8319	0.8052	0.7962
Stock Only	0.5619	0.3157	0.5619	0.4043
Article Only	0.8044	0.8307	0.8044	0.7953
Model-Predicted Credibility	0.7959	0.8046	0.7959	0.7904
Domain-Based Credibility	0.8025	0.8296	0.8025	0.7931

Table 3: Performance metrics of stock movement prediction models.

1	China's large manufacturers are gaining size through overseas deals
2	Trump orders mental health support for military veterans to prevent suicide
3	This investor is still bullish on tech stocks in China
4	BioMarin CEO on hemophilia treatment success
5	Celgene chief Alles discusses deal to acquire Impact Biomedicines, strategy for 2018

Table 4: Titles of articles predicted correctly without credibility and incorrectly with credibility.

assignments.

7 Ethical Considerations

One thing that we must take into consideration is that our data consists of articles only in English. This alienates the majority of the world and only analyzes how English speakers influence the S&P500 Index. Additionally, because the S&P500 index only consists of U.S.-based companies, the results and models may be unable to be generalized to the entire stock market, but may only apply to companies in the United States. We must also consider the data bias in news articles which may lead the model to portray certain demographics as more influential on the index price than other groups.

Another important thing to remember in using these models is that the S&P 500 is an index of the 500 companies in the United States that have the highest market cap, and using this model to invest in the index is no different than to invest in these companies. As profitability often does not coincide with moral virtue, users of this model are encouraged to consider the effects of investing in this index fund.

8 Work Distribution

Brian Lucas - led the major work on the stock price prediction model and also contributed in the Dataset creation

Owen Wurst - led the major work on the credibility model and also contributed to the literature review

Kasvitha - led the major work on the literature review and also contributed on the credibility model

Kartik Hans - led the major work in the Dataset creation and also contributed on the baseline model

References

[1] Gidofalvi, Gyozo. (2001). "Using News Articles to Predict Stock Price Movements". *UCSD*, https://people.kth.se/g̃yozo/docs/financial-prediction.pdf.

[2] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu. (2019). "Stock Price Prediction Using News Sentiment Analysis". *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 205-208.

[3] Li, Y., Pan, Y. A novel ensemble deep learning model for stock prediction based on stock prices and news. Int J Data Sci Anal 13, 139–149 (2022). https://doi.org/10.1007/s41060-021-00279-9

[4] "Stock Movement Prediction from Tweets and Historical Prices", Yumo Xu and Shay B. Cohen

[5] "Stock Price Prediction Using BERT and Word2Vec Sentiment Analysis", Willem Sebastian and Sani M, Isa

[6] S. Kogan, T. Moskowitz, M. Niessner, "Fake News in Financial Markets," 2021, https://spinup-000d1a-wp-offload-media.s3.amazonaws.com/faculty/wp-content/uploads/sites/3/2021/08/Fraudulent-News-Direct-and-Indirect-Effects-of-Financial-News-Manipulation.pdf.

[7] Evans, Lewis, et al. "Credibility assessment of Financial Stock Tweets." Expert Systems with Applications, vol. 168, Apr. 2021, p. 114351, https://doi.org/10.1016/j.eswa.2020.114351.

[8] Shahane, Saurabh. “Fake News Classification.” Kaggle, 8 Oct. 2023, www.kaggle.com/datasets/saurabhshahane/fake-news-classification.

[9] Sumanthvrao. "Fake-News-Dataset." Kaggle, 19 Apr. 2019, www.kaggle.com/datasets/sumanthvrao/fakenewsdataset/data.

[10] Lin, H., Lasser, J., Lewandowsky, S., Cole, R., Gully, A., Rand, D. G., & Pennycook, G. (2023, September 2). High level of correspondence across different news domain quality rating sets. OUP Academic. https://academic.oup.com/pnasnexus/article/2/9/pgad286/7258994?login=false

[11] Jeet.J. (2018, September 5). US financial news articles. Kaggle. https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles

For all of the code and data for this project, go to https://github.com/OWurst/PittCoin