The Impact of Varied News Sources on FinBERT-based Stock Predictions
Owen Bluman

With just a surface level knowledge of financial markets and endless free time during the pandemic, there was no better time for me to get into the stock market than the fall of 2020. The scrapped together remains of leftover birthday money came together to form my first ever investment account on the Robinhood platform. Following the advice of my parents I had accumulated a diversified portfolio of low risk index funds and now I couldn't wait for the money to start rolling in. The only problem was that it was now 4 months later and my account balance was still only slightly higher than when I began. This was a far cry from the endless stream of cash I had envisioned. Naturally, I then decided to take advice from a new source, one filled with big dreams and bigger losses, r/wallstreetbets.

After seeing a flood of posts championing GameStop as the next big riser, I threw my entire portfolio into the stock and hung on for a rapid ascent, followed shortly by an unrelenting nosedive (Figure 1). With almost nothing left, I returned to the poster child of traditional investment media, the trusty S&P 500 and patiently endured almost 1.5 years until my account returned to its initial value. This experience has left me with a lasting motivation to explore how accurately different sources of financial news can predict the future of the stock market, in order to determine for myself who to trust for investment advice.

Figure 1: Graph of authors Robinhood account value from Fall of 2020 to Spring of 2022

The goal of this audit is to fulfill that very impulse, and determine which sites offer the most sound stock advice. Specifically, I would like to collect headlines from a variety of traditional online news sources, such CNBC.com, CNN.com, and Forbes.com, as well as post titles from the r/wallstreetbets, r/investing, and r/stocks subreddits of the alternative social media platform Reddit. After collecting this dataset of market information, I want to use the trends

drawn from these sources to predict the movement of both the S&P index fund as well as individual stocks including Tesla, Nvidia, and Google, and compare the accuracy of forecasts derived from each news source. However, getting meaningful insight from a sea of hundreds of headlines and post titles is not an easy task. Fortunately the evolving field of natural language processing provides an answer in the form of artificial intelligence systems trained for sentiment analysis.

The FinBERT model is one such system that is specifically tuned for sentiment analysis of financial text. Built as a collaboration between researchers Allen H. Huang and Yi Yang from the Hong Kong University of Science and Technology and Hui Wang of the Renmin University of China[1], the model is an extension of Google's BERT language model. Huang et al. collected a large corpus of financial data including over 200,000 corporate filings (10-Ks and 10-Qs), nearly half a million financial analyst reports, and more than 130,000 earnings call transcripts in order to use for pre-training on the open source BERT model (Figure 2). Then they fine-tuned the FinBERT on an annotated financial sentiment dataset, while implementing a final softmax output layer with three sentiment class labels (positive, negative or neutral).

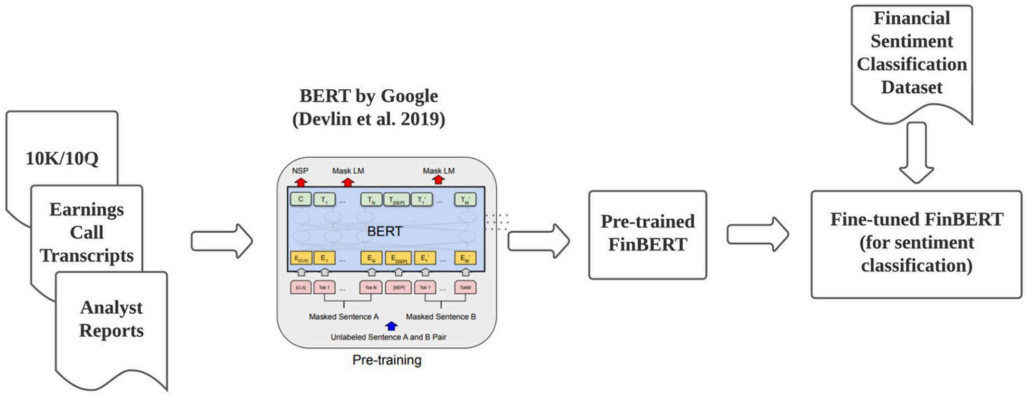

Figure 2: Diagram of FinBERT model training architecture

Given FinBERT's design and the established financial data used to train the model, I hypothesize that the traditional news sources will be more accurate at predicting the movement of individual stocks. This is due to my belief that the trends of specific companies are more likely to be intercepted by experts with specialized knowledge and ties to specific industries, and these experts are more commonly found working for a traditional news source. Additionally, I hypothesize that Reddit will be more accurate at predicting the overall market (S&P 500) as the Reddit dataset will allow a larger and more representative group of investors to sway the final prediction due to the lower barrier to entry to post on a subreddit, versus publishing a news article for a large company.

The data used for the audit, news headlines and Reddit post titles, was appropriate and even ideal for the previously stated goal of comparing market predictions gleaned from different financial sources as headlines and titles serve as the sources most visible pieces of information that a typical user would be more likely interact to see and in turn, make decisions from. Given that the average user won't click on every single article or post, it can be assumed that the headlines and titles will be viewed much more often than the actual content of the piece of media itself.

The specific data identified from the traditional news sources come from three of the largest traditional financial news sources based on reputation, and the the three largest financial subreddits based on user count, thus making it likely that a standard investor has interacted with these sites before and would make decisions based on their information. For the purposes of the audit, headlines and posts published in the week leading up to the prediction date (the week preceding April 21st) were considered to respect the timeliness of financial information. The specific individual stocks and their accompanying datasets were chosen to reflect their prevalence in discussion, as an investor would presumably be more likely to trade stocks with a higher volume of information regarding them circulating in their news bubble.

The traditional news headlines and the reddit post title datasets were collected via separate methods. The headlines were scrapped via the "News API", which returned headlines given an internal website search queries for the terms "market" as well as the individual stock companies names themselves, "Tesla", "Nvidia", and "Google". The data for individual stocks were then further filtered out to only include titles with an explicit reference to the company name or stock ticker, as a backup check to make sure the queries only returned relevant information. Finally, the results for each term were stored in a corresponding CSV file, with each line being a separate headline. The Reddit post titles were scraped in a similar fashion, but using the official "Reddit API", which simply returned all recent posts. The same filtering process was done to only include posts with an explicit mention of "market" or the company name, or additionally the stock ticker for each security (SPY in the case of the overall market as it saw a higher volume of information than the direct S&P). These results were stored in separate individual CSV files in the same fashion.

With the dataset collected, the pipeline for the audit of FinBERT's sentiment analysis ability on varied datasets can be established. The audit specifically loaded the finbert-tone version of the model from the freely available Hugging Face Transformers library. Given the loaded model, each of the CSV files corresponding to a given stock (S&P in the case of the overall market) is input line by line to FinBERT which outputs a sentiment class prediction (positive, negative, or neutral) and a confidence score in the predicted class. The confidence scores for each predicted class are accumulated for each file, in order to appropriately weight headlines which provide stronger sentiment. This method is designed to counter headlines that

might be more ambiguous worded and may result in the model being less confident in the predicted class. Each source is then assigned a net sentiment score for a given stock, normalizing for the given number of relevant headlines, calculated by the equation:

$$Net\ Sentiment\ =\ (Total\ Positive\ Score\ -\ Total\ Negative\ Score)\ /\ Total\ File\ Headlines$$

Next, the closing price data for each stock for each day in the week leading up to the prediction (data for April 14-17th) was manually collected and hardcoded for the purposes of the audit. From this data, the average stock movement per day can be calculated by a simple average of the differences in between each day's true price, in order to determine how relatively volatile the stock is over the given time period to be audited. Lastly, the final predicted price was calculated using the formula:

$$Predicted\ Price\ =\ Previous\ Closing\ Price\ +\ (Net\ Sentiment\ *\ Average\ Stock\ Movement\ *\ Scaling\ Factor)$$

This equation effectively takes the final price of the stock on the open market during the period when the headline data is being collected, then predicts the direction and magnitude of a change in the stocks price based upon the net sentiment analysis derived from FinBERT's analysis, weighing the average amount in which the stock has been moving in the past week, and then scaling appropriately to make sure the predictions have a defined impact. The scaling factor was arbitrarily set at 10 for the individual stocks in order to visually show predicted impact, and then for the S&P 500 prediction the scaling factor was relatively set at 20 to account for the much greater stock price of the index fund in comparison to the individual stocks, and thus a necessary higher scaling factor.

The results of the audit were graphically presented as a line graph of the true stock price through the week preceding the prediction date and up to and including the prediction date itself, surrounded by dots representing the predicted prices on the prediction date, colored according to the traditional news source it corresponds to. Additionally, the metrics of signed percentage difference between the true and predicted prices, and the average absolute percentage difference for each classification of news source were calculated and output in a text file. The first statistic serves to show how close each news source was to the true price and whether the estimate was overly optimistic or pessimistic, while the latter compares the average accuracy of traditional news sources against the Reddit sources for the specified stock.

Some findings of note include the graph of the S&P 500 predictions (Figure 3) which visually show the most accurate prediction of the entire audit, as the prediction of the stock price calculated from the CNBC sentiment analysis was off by less than two dollars, for a signed percentage difference of only 0.03%.

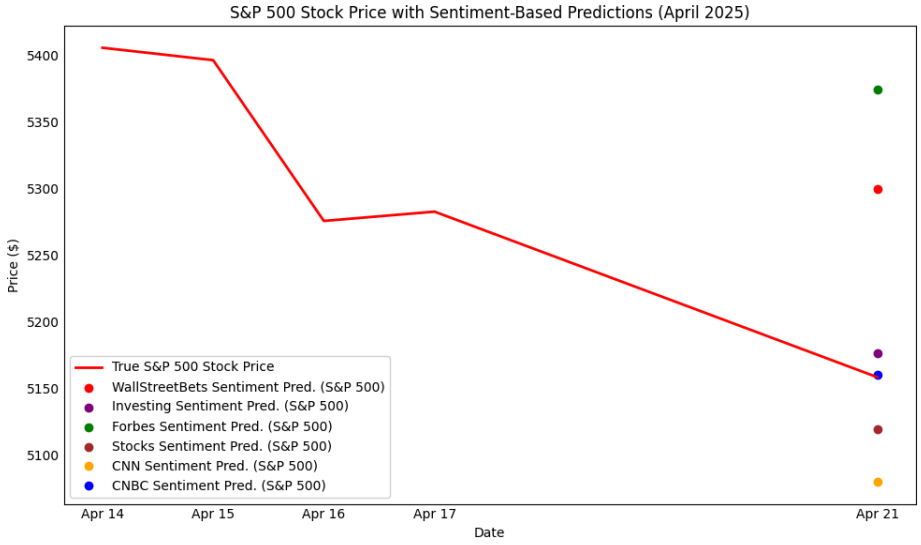

Figure 3: Graph of S&P 500 true stock price for the week leading up to and including the prediction date, alongside each source's predicted price

This chart also confirms my second hypothesis, that Reddit would be more accurate at prediction of overall market trends as the Reddit sources for the S&P had an average absolute difference of 1.28% compared to the Traditional Sources result of 1.92%. This can be seen visually by the r/investing" and r/stocks subreddit predictions being much closer to the true line than that of Forbes or CNN.

In contrast with some of the relatively close predictions of the S&P 500 compared to the stocks true value found above, the predictions for Nvidia were much less accurate (Figure 4). Both the traditional and alternative sources had their highest average absolute difference out of any of the stocks of 3.8% and 8.31% respectively (Figure 5), and no individual source was within 2% of the true value. The r/stocks based prediction was notably inaccurate as it predicted a stock value 15.15% lower than the true value, the largest difference between any individual prediction and any true stock value.

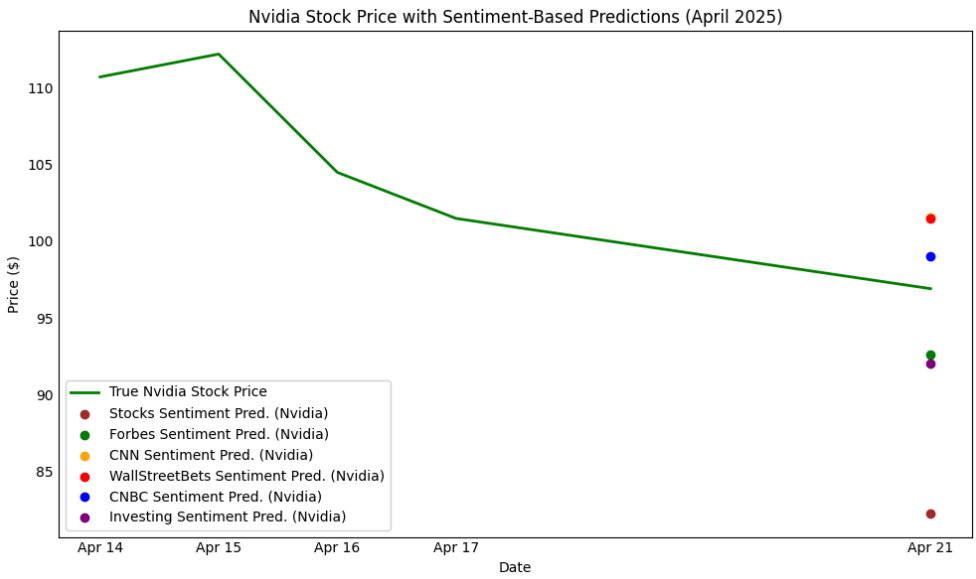

Figure 4: Graph of Nvidia true stock price for the week leading up to and including the prediction date, alongside each source's predicted price

```
Nvidia Stock Price Predictions - April 21, 2025
True Price: $96.91

Predicted Prices by Source:
Stocks: $82.23 (-15.15%)
Forbes: $92.63 (-4.42%)
CNN: $101.52 (+4.76%)
WallStreetBets: $101.49 (+4.73%)
CNBC: $99.06 (+2.21%)
Investing: $92.02 (-5.04%)

Average Absolute % Difference (Traditional Sources): 3.80%
Average Absolute % Difference (Reddit Sources): 8.31%
```

Figure 5: Text file output of statistics calculated from Nvidia stock price predictions

Another note is that the CNN prediction dot cannot be seen very well on the graph due to its extreme similarity to the r/wallstreetbets prediction, as the predicted stock values were only three cents apart. The breakdown of the average absolute difference plays a role in establishing the truth of my first hypothesis, that analysis of traditional news sources would perform better on individual stocks then on Reddit posts as the traditional source had 4.51% less average absolute error. This was further confirmed by the results of the Google and Tesla predictions, as the average absolute difference of the traditional sources were lower than that of the subreddits on these individual stocks as well.

Other more general findings of interest include r/wallstreetbets propensity to overestimate a stocks value, as the prediction gleaned from its data had a positive signed difference above 2% for three out of the four stocks, which aligns with my experience detailed in this paper's motivation. Furthermore, CNBC was the most accurate predictor of stocks overall as it had the closest prediction for both the S&P 500 and Nvidia, the second closest for Google, and nearly tied for the second closest for Tesla by a margin of 0.01%.

The above results and findings provoke a number of interesting societal implications about the way in which machine learning models perform. Given that my hypothesis that individual stocks were more likely to be accurately predicted with FinBERT's analysis of traditional media as opposed to alternatives turned out to be correct, it calls into question the power imbalance between those who create models and those who use them. Since data from traditional sources can sometimes be inaccessible, requiring paid subscriptions or costly API queries, if companies decide to develop models for public use that perform better when fine-tuned with proprietary data, the issue of fairness for the user is brought to light. One can view a user's ability to access such data for use in training can be likened to a protected feature being used in a classification algorithm.

The question of whether the model should provide different levels of accuracy based on a user's ability to afford to supply it with certain background material is one that may not have an obvious answer. Damping a model's ability to learn patterns from a specific group of data in an equitable manner for instance would level the playing field in terms of the models performance given different pre-training processes, but it may hurt the models performance overall. As stated in *Big Data's Disparate Impact* by Barocas and Selbst, "attempts to ensure procedural fairness by excluding certain criteria from consideration may conflict with the imperative to ensure accurate determinations[2]." The choice to change the model's architecture to account for certaining training biases could certainly have an impact on the way the model is currently being implemented. Even this decision is taken out of the users control, as the ways in which potential fairness measures are implemented are created without the input of those relying on the system itself.

Furthermore, the existing corpus of financial data used in FinBERT's training may exhibit traces of bias as well. If the more subjective assets used to train the model contained unfavorable bias against a certain company or industry, then the sentiment analysis results based upon text related to these tainted groups have the potential to be skewed negatively. To again impart from *Big Data's Disparate Impact*, "data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society[3]." While documents such as corporate filings are for the most part purely fact based, the analyst reports and earnings call transcripts that make up a substantial

portion of FinBERT's training data are certainly infused with human bias. If in the future FinBERT or a similar implementation are used to make real trades, this could start a feedback loop of discrimination, as models stray away from a certain classification of investments which in turn lower the value of the investment, further decreasing the chance that a model might choose to invest in it.

Despite the potential bias contained within its training data the researchers who developed FinBERT did make the honorable decision to publicly release the exact sources from which they got the data to fine-tune the model. This is a choice that large commercial models such as the latest versions of GPT, Llama, and Gemini have not explicitly followed. While it can open the door to discriminatory output and bad press if there does turn out to be issues with the way data is collected or conflation between features and protected classes, open access to training data is a necessary objective in order to hold model creators accountable. As Barocas and Selbst stated in *The Intuitive Appeal of Explainable Machines* "subjective decisions pervade the modeling and decision-making process. Explaining why the model works as it does requires accounting for these decisions.[4]"

Those creating models should certainly have to face repercussions if their design choices are the reason why a model exhibits certain biases. Barocas and Selbst place great emphasis on understanding why and how non intuitive hidden patterns are revealed in the training process and FinBERT's authors took a step on the right path with their choice. Additionally, this decision can lead to increased interpretability for the average user, and they may be able to derive their own reasoning for why a model outputs a particular value given a certain input.

This methodology of this audit has a few limitations that are necessary to critique. First, the time scale of the audit is relatively short, as data is collected and predictions are made all within the span of a single week. Furthermore, the scope of the audit is capped to four different stocks, which represents an incredibly small portion of the overall trading volume on any given day. This of course opens the possibility that the trends found by this sample of the market are not generalizable to the average experience of an investor. Additionally, while headlines and post titles often contain the most widespread information that any given user is more likely to encounter, choosing these data points as the sole source of financial predictions may leave out more deep thought predictions that are hidden within the overall article.

In the context of this audit, FinBERT exists in a vacuum, and other financial sentiment analysis models could have been used to provide a baseline or comparison of results, in the case that FinBERT's implementation specifically performs worse in certain situations. The price prediction equation also contains a possible point of objection, as the scaling factor used to enhance the significance of the model's predictions may also destroy the underlying patterns that are contained in the data.

To complement the audits methodology, the choice of sources are extremely relevant, high trafficked sites that generate a large amount of financial content, making them ideal to collect data from. Furthermore, weighting the confidence score of the predicted class into the final equation tried to gain more insight into FinBERT's full output, and hopefully resulted in more accurate predictions overall. To add, the acknowledgment of the stocks volatility in the recent past is a metric that could have some positive influence on the ending accuracy, even though this metric may be influenced by a variety of outside factors such as upcoming earnings or dividends, and public events influential to the companies financial performance.

The findings of the audit also deserve scrutiny, as there is a chance that correlations between sentiment and stock price are coincidental and unrelated. An even further potential event is that there is reverse causation where changes in stock price are actually the events altering the discussions on the news sites and subreddits. I am inclined to believe that both financial analysts and opinionated Redditors would have enough conviction and confidence to make true predictions and not just tail the market trends.

I equally believe that there are positive critiques of the audit results, as the charts and statistics used are relatively straightforward and intuitive. Furthermore, the results directly allowed the author to test and evaluate the accuracy of both hypotheses, a necessary and sometimes overlooked part of the audit process. I also believe that the audit is straightforward to repeat and extend, given the lack of difficult mathematical notation contained and these easy of loading and utilizing the model in the provided code

As delivered by Danaë Metaxa et al through *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In*, audits are activist and "the end goal should be to draw attention to an issue and bringing about political and social change[5]." This audit has certainly prompted a personal examination of the source and type of news I regularly receive, as well as the news I hope to consume in the future. The era of information being shared freely and publicly by any number of diverse users has already begun, and based on the results of this audit, alternative news sharing platforms may even overtake traditional news sources. If Reddit can have the potential to usurp industry news standards in an industry as consequential as investing then any number of fields may be next.

To build upon the previous point, there are several extensions of future work that I can foresee this audit establishing, first of which being a political analysis. I would like to examine on a political axis (Figure 6) how different news sources have their sentiment analyzed, particularly when covering the same topic. This could include a comparison of their economic reporting by utilizing FinBERT again, or it could come in the form of another sentiment analysis model tuned for political sentiment.

Figure 6: Media bias chart depicting the categorized political affiliation of major news outlets, a potentially in use for an extension of this audit

A more direct successor to this audit could be a similar dataset pipeline but instead with a more advanced price prediction model, perhaps using a time-series forecasting algorithm in conjunction with the existing sentiment analysis. There are also options such as boosting or RNN's that might provide an interesting comparison of pricing accuracy among models.

In final analysis, auditing FinBERT has proved a worthwhile and fulfilling endeavor. I am proud to contribute to the greater artificial intelligence community by examining how a sentiment analysis model responds to different types of input, while examining my own methodology and findings in the process. While the sample size of my investigation may not be fully conclusive, it seems that perhaps listening to r/wallstreetbets for stock advice really was a mistake all along.

Endnotes:

1. Allen H. Huang et al, “FinBERT: A Large Language Model for Extracting Information from Financial Text” 813.
2. Solon Barocas and Andrew D. Selbst, “Big Data’s Disparate Impact.” 722.
3. Barocas and Selbst, “Big Data’s Disparate Impact.” 674.
4. Andrew D Selbst and Solon Barocas, “The Intuitive Appeal of Explainable Machines.” 1129-1130.
5. Danaë Metaxa et al, “Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.” 322.

Bibliography

Barocas, Solon, and Andrew D. Selbst. “Big Data’s Disparate Impact.” *California Law Review*, vol. 104, no. 3, 2016, pp. 671–732, papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899, https://doi.org/10.2139/ssrn.2477899.

Huang, Allen H., et al. “FinBERT: A Large Language Model for Extracting Information from Financial Text†.” *Contemporary Accounting Research*, vol. 40, no. 2, 29 Sept. 2022, https://doi.org/10.1111/1911-3846.12832.

Metaxa, Danaë, et al. “Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.” *Foundations and Trends® in Human–Computer Interaction*, vol. 14, no. 4, 2021, pp. 272–344, https://doi.org/10.1561/1100000083.

Selbst, Andrew D., and Solon Barocas. “The Intuitive Appeal of Explainable Machines.” *SSRN Electronic Journal*, vol. 87, no. 3, 2018, https://doi.org/10.2139/ssrn.3126971.