Reading the Market

Sentiment Analysis of Financial News Data for Stock Price Prediction

Connor Malley
malleyconnor@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

ABSTRACT

In this paper, I analyze a variety of sentiment analysis models for the forecasting of prices in the stock market. Many investors today believe in the efficient market hypothesis[5], which states that the stock market is efficient in the sense that the current security prices incorporate all past information. However, many investors have been able to outperform the market as a whole over long periods of time using technical indicators, fundamental analysis, valuation, analysis of news data, and extensive industry knowledge. Therefore, I make use of news data to see if it is possible to predict the same-day performance of individual stocks/indexes with better than random results. I also compare pre-trained BERT models to a baseline model to examine if training on highly specific financial data has any benefit in terms of the accuracy.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Information extraction**;

KEYWORDS

sentiment analysis, natural language, news datasets, market forecasting

ACM Reference Format:
Connor Malley. 2018. Sentiment Analysis of News Data for Stock Price Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Forecasting the stock market has always been a problem which has captivated data scientists, mathematicians, and deep learning engineers alike. According to the Efficient Market Hypothesis (EMH), the stock market is efficient and a securities current price should reflect all prior information. [5] However, many investors, such as the famous Warren Buffet, have been able to historically outperform the market. This has been done by using multiple modes of information such as news data, company financial data, technical indicators in stock prices, and even insider knowledge which is not immediately available to the public. The father of the efficient market hypothesis, also showed strong evidence that historical stock prices display random walk patterns, and thus future stock prices should be virtually impossible to predict with previous price data. [4] Thus, that should leave either news data, financial data, industry knowledge, and insider knowledge as the only alternative indicators for a stocks future price. As ready access to the latter three is not feasible on a large scale for deep learning, I seek to train a variety of natural language deep learning models on datasets of historical news data. First, I try training an LSTM using pre-trained word embeddings to predict the sentiment, or direction, of the DOW Jones Industrial Average (DJI). Then, I try predicting the direction of individual stocks using a larger dataset of specific financial data from investing.com using a pre-trained Bidirectional Encoder Representation of Transformers (BERT) model. Finally, I try the predicting the direction of individual stocks on a much larger dataset of analyst ratings from benzinga.com. I also conduct some ablation studies on the models and look at the effects of finetuning, vs. predicting with pre-trained sentiments without any additional training.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 PROBLEM STATEMENT

Predict the direction of stock price using a variety of deep learning language models on news headlines, and examine how the sentiment of news headlines relate to stock price.

3 RELATED WORK

It was shown that there is relation between stock news data and the stock stock price. (Khedr et. al, 2017) shows that the movement of a stocks price can be predicted with 59.18% accuracy using a K-Nearest Neighbors Classifier on stock news sentiment from a Naive Bayes classifier. [7] (Shah et. Al, 2018) develop a dictionary-based sentiment analysis model which was able to display 70.59% accuracy in predicting the short term direction of the stock market.[11] (Attigeri et. al, 2015) tries forecasting directions of stock prices using a sum of word sentiments in news data. [2]

4 TECHNIQUE & EVALUATION

4.1 Predicting Dow Jones Performance with World News Data

I first tried to utilize a dataset from Kaggle, which consists of 8 years worth of daily world news headlines from Reddit [12]. This dataset has 25 headlines per day, covering 1989 days, all pulled from the top posts on the Reddit channel r\worldnews channel from the years 2008 - 2016. This dataset has a binary label of whether the Dow Jones Industrial Average(DJI) went up or down on each given day. The task is then to predict the direction of this stock index on a given day, by processing the text data in some manner. One intuitive approach for processing this data is by training a Long Short Term Memory (LSTM) model on the data, as well as

an embedding layer. Of course there is also the option of using pre-trained embeddings trained on a much larger dataset, such as the GloVe 6B embeddings[10], which were trained on the Wikipedia 2014 and Gigaword 5 datasets. In this paper I opt to experiment with the pretrained GloVe embeddings given the small size of the dataset. First, I start by expanding the number of samples in the dataset by splitting each day into 5 different chunks, so that only 5 headlines will be used for each prediction. This creates a dataset of 9945 samples, compared to the original 1989 samples. For each sample of 5 headlines, each headline is passed through an LSTM as a "mini-batch", and the final hidden state of each headline is concatenated and passed through a fully-connected layer for classification, as detailed in Fig 1.

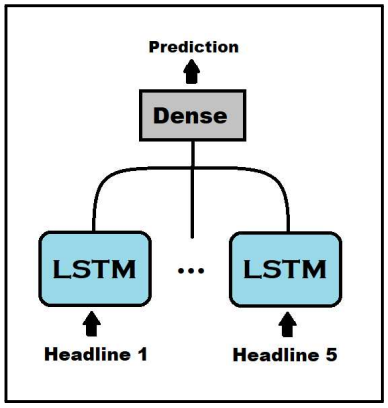

Figure 1: Reddit News LSTM Model

4.1.1 Implementation Details. For the LSTM model I used a 2-layer bidirectional LSTM with a hidden size of 64, and a single linear layer with input size 2*64*5 and output size 1 for the sentiment. I trained this model for 8 epochs using the Adam optimizer with a learning rate of 5e-3 and batch size of 4. To help the validation loss converge, I used a dropout rate of 0.3. For the embeddings, I represent each word using the pre-trained GloVe 6B 100-dimensional embeddings [10].

4.1.2 Results. After extensive testing and experimentation with the hyperparameters, the validation loss was unstable and did not converge while the model started to memorize the training dataset. However, the accuracy on the validation dataset does stay slightly better than random, and fluctuates while the training accuracy continues to increase, as shown in Figure 2. During some epochs the validation accuracy reached up to 57%, however because of the stability of the model, these results were not easily repeatable.

This clearly indicates that either there is not enough data in the training dataset, which is likely given the extremely small sample size, that there is no strong correlation between the world news data from Reddit and the Dow Jones performance, or that more regularization techniques need to be employed to achieve stable convergence of the validation loss. However, even after testing with higher dropout rates up to 0.5, and smaller hidden sizes down to 16, the validation loss continued to remain unstable and not much better than random. In many cases, the model simply degenerated

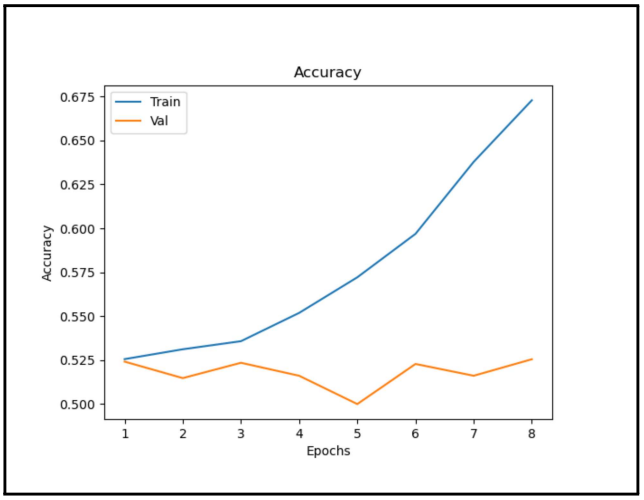

Figure 2: Reddit News LSTM Model Accuracy

to always predicting a positive sentiment the large majority of the time, even with the incorporation of weight decay. I also want to note that even the top Kaggle notebooks utilizing various methods such as LSTM, MLP with a TFIDF vectorizer, and SVM were not able to achieve much better results.

4.2 Predicting Stock Performance with Financial News Data

The Reddit world news dataset is relatively small in size, and I could not find any other attempts that achieved stable, repeatable results. Also, many of the top news headlines obviously had nothing to do with the performance of the American stock market (e.g. "Courts put 13yearold girl in state care blocking her from being the youngest person to sail around the world"). Therefore, I downloaded another dataset which consisted of historical financial news data relating to 800+ stocks from investing.com. [6] Howvever, this dataset did not have any labels, so I used the Yahoo Finance Python API to label each news headline with whether the stock went up or down that day.

In this dataset, I only consider those 505 stocks which are currently in the Standard and Poor's 500 Index (S&P500), which leaves me with 123858 news articles from October 2008 - May 2019 in the training set, and 32705 news articles from May 2019 - February 2020 in the validation set. These news articles are all highly specialized financial data, unlike the general news data in the Reddit dataset, and thus using a pre-trained sentiment model on general news or ratings data (such as the IMDB movie review dataset) [8] to generate the predictions will likely not produce good results. This is partly because many of the tokens present in financial news (such as stock tickers and other finance-specific terminology) are not present in many sentiment analysis datasets. Therefore, I opted to use a pre-trained BERT model, which was pre-trained using the Masked Language Model (MLM) and Next-Sentence Prediction tasks. This model, dubbed as FinBERT [1], was first pre-trained on the BookCorpus+Wikipedia dataset[13], then further pre-trained on the Reuters TRC2-financial dataset, which is a subset of size

46143 financial news documents from the original TRC2 dataset.The model was then fine-tuned for sentiment analysis on the Financial Phrasebank [9] dataset, a dataset of 4837 sentences from financial news articles labeled with positive, negative, or neutral sentiment by 5-8 annotators. The model consists of 12 transformer encoder layers, 12 attention heads, and a hidden size of 768, where the sentiment predictions are generated using an appended classification token.

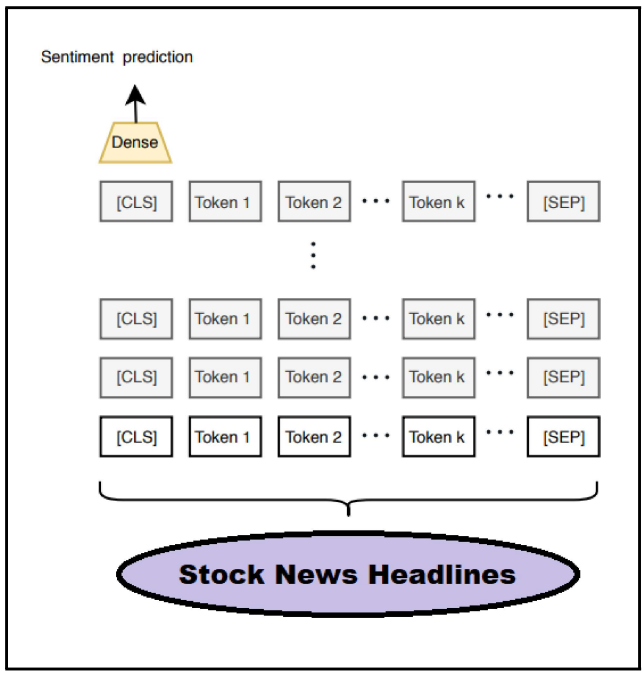

Figure 3: FinBERT Model

This model, which is shown in Figure 3, must be altered to produce positive and negative sentiments only since there is no neutral sentiment for the dataset which I've created. Therefore, when fine-tuning on the stock news dataset, I swap out the classification head for another one which generates two outputs in the final linear layer, and train it from scratch. To speed up the training process, I also start by freezing every layer of the FinBERT model, except for the final 3 encoder layers and the classification head as detailed in [1], as this method of training was still able to achieve good results on the Financial Phrasebank dataset with a lower training time.

4.2.1 Implementation Details. When fine-tuning, I used 0.1 Dropout and ReLU activation on the output of the encoder, and a fully-connected neural network with 1 hidden layer of size 768, mapping the 768 output features to 1 sentiment score, which is scaled to $[0, 1]$ with Sigmoid activation. I fine-tuned the model for 8 epochs with a learning rate of $1e-5$ using the Adam optimizer. As a comparison, I examine the results if I were to just predict the direction of the same-day stock movement using the pre-trained sentiment. Again, since the pre-trained sentiment has an additional logit corresponding to neutral sentiment, I just take the argmax of the positive and negative logits to get a final prediction. Also, when pre-processing the data, I replace every occurrence of the respective stock ticker with a [MASK] token to prevent possible splitting by the tokenizer. If the stock ticker is present in the sentence, the model should also have access to additional context present in the news headline (e.g. "[MASK] went up today, while AAPL went down").

4.2.2 Results. For the baseline of directly predicting the price direction using the pre-trained sentiments of the news headlines without any fine-tuning, the model achieves 52.2% accuracy on the validation set. This is slightly better than random, indicating that there is little correlation between the zero-shot sentiment and the direction of a stocks price in the day the news headline was published.These results can be visualized in Figure 4 I found additional fine-tuning of the model to directly predict the movement of the stock price to provide very little additional benefit, if any, as the training process is again very unstable. This model achieved 52.8% accuracy with additional finetuning, only a 0.6% increase from the baseline zero-shot model. The training accuracy again continued to increase while the validation accuracy remained not much better than random. This could again be an effect of many different factors, such as the size of the dataset, hyperparameters of the model such as as the learning rate or dropout rate, or even the data preprocessing using the pre-trained tokenizer. It is also a question of whether the news data in this dataset has any relation to the same-day stock price. After examining the training dataset, I found that many of these article headlines have little to relevance to the associated stock. For example, the article headline "After hours Gainers/Losers" provides no information as to whether the stock in question, NVDA, was a gainer or loser for that day. Also, for many datapoints, the stocks associated with the article headline were not the main subject of the article. This brings into question the validity of the data-gathering and labeling process for this dataset, thus in a final effort to show some decent results I searched for another much larger and more reliable dataset to test on.

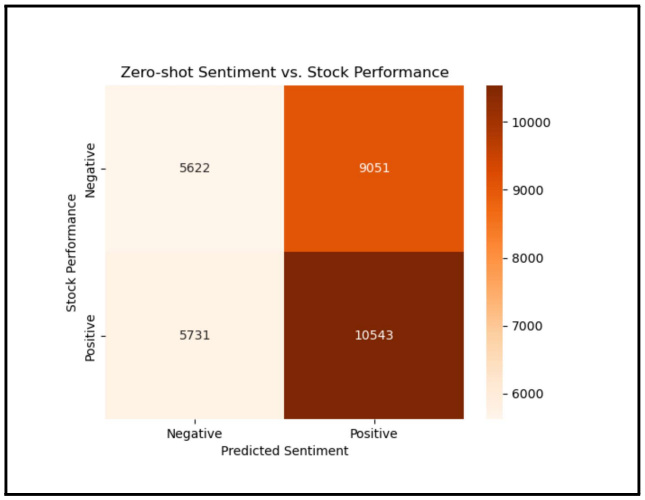

Figure 4: Confusion Matrix for Baseline Sentiment Prediction on Financial News Dataset

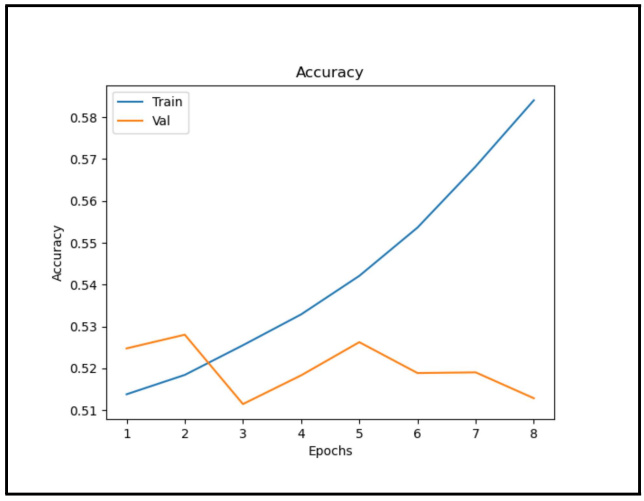

Figure 5: Results for FinBERT Finetuned on Financial News Dataset

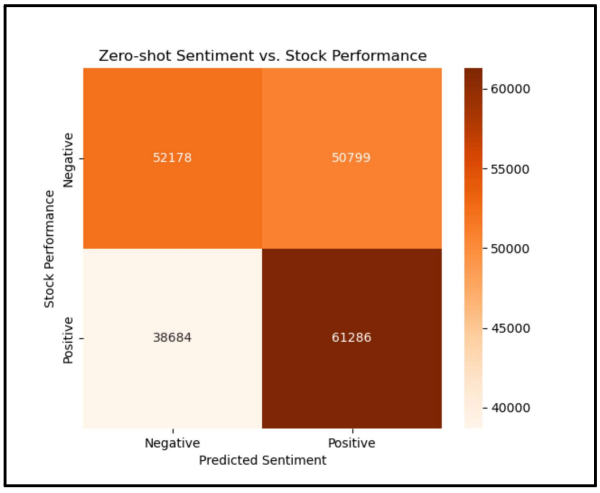

Figure 6: Confusion Matrix for Baseline Sentiment Prediction on Analyst Ratings Dataset

4.3 Predicting Stock Performance with Analyst Ratings Data

This dataset used in this section consists of analyst ratings from benzinga.com for 6000+ stocks over a period of 11 years. [3] This dataset has a comparatively larger size of 811784 headlines in the training set and 202947 headlines in the validation set. Obviously, if the amount of data is of any concern for this prediction task, there should be an increase in performance when training on this dataset.

4.3.1 Implementation Details. For the implementation with this dataset, I use the similar hyperparemeters for FinBERT as with the previous dataset. This corresponds to a 0.2 dropout with ReLU activation in the hidden layers, and a sigmoid activation in the final layer. I also use a learning rate of $1e-5$ in this experiment.

4.3.2 Results. Here I found similar results as with the previous two datasets, wherein the training accuracy continues to increase while the validation accuracy fluctuates around the same value. Just looking at the confusion matrix for the zero-shot sentiment baseline, it is clear that this dataset has much higher correlation between the predicted sentiment of the news headlines and the corresponding price direction of the stock.

However, the validation accuracy is 57.4% for this dataset, which is much higher than for the previous two datasets. I believe this is mostly a result of the dataset itself, as this dataset has a much greater relation between the predicted sentiment of the news headline and the direction of the stock price. Looking at figure 6, it is shown that if I use the zero-shot sentiments of FinBERT to predict to stock direction, this results in an accuracy of 55.9%. Thus, with additional finetuning, the FinBERT model was only able to squeeze out an additional 1.5% in accuracy for predicting the stock price.

Obviously, even with a dataset as large as this, these results show that the size of the dataset may not be the issue, but rather the relevance of the data itself as well as the correlation between the sentiment of news on a stock and the respective stocks price.

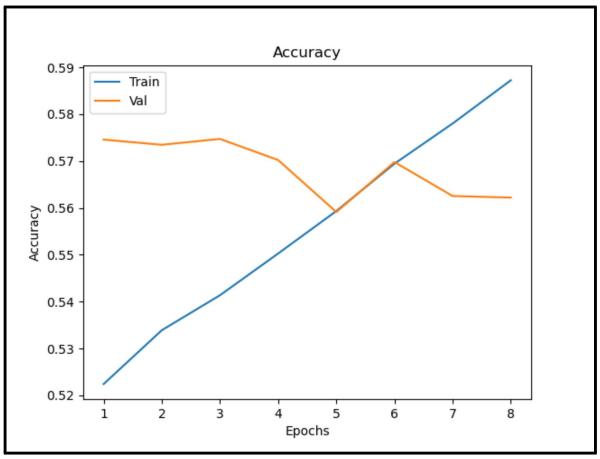

Figure 7: Results for FinBERT Finetuned on Analyst Ratings Dataset

Comparing figure 8 with figure 6, it is clear that the model is learning some new information which is able to predict the stocks price better than just the sentiment of the news headline alone. After examining these results, it may be that a stocks same-day price can only be predicted with so much accuracy using a data sample such as a news headline. I believe that even more data may be useful for training, so it may be worth examining larger datasets which may have samples on the order of billions, or even datasets with synthetically created sentences with either positive or negative sentiments for some stock.

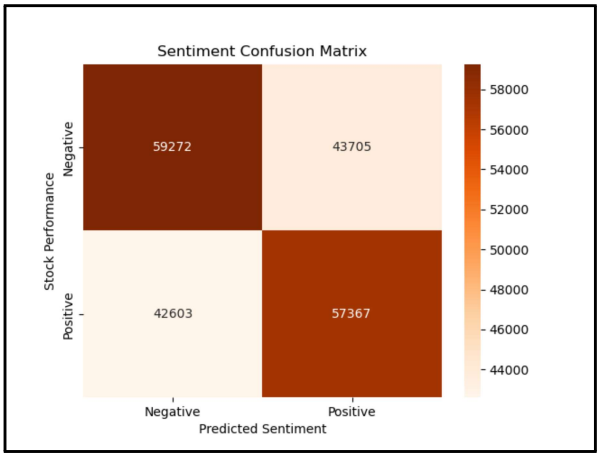

Figure 8: Confusion Matrix for FinBERT Finetuned on Analyst Ratings Dataset

4.4 Performance with Different Price Labeling

I wanted to examine if it was possible to predict longer range dependencies of Analyst Ratings Data in stock price difference. Therefore, I created three new datasets based on the original Analyst Ratings dataset. The first dataset is just the original dataset. The 2nd dataset was labeled with the next-day price difference for each stock, where if t is the day that the article was released, $label = (close(t+1) - close(t) > 0)$ where $close(t)$ indicates the closing price of stock on day t. The 3rd dataset was labeled with the next-five day price difference for each stock, where $label = (close(t+5) - close(t) > 0)$. Lastly, the 4th dataset was labeled with the next-twenty day price difference for each stock, where $label = (close(t+20) - close(t) > 0)$. Each dataset has approximately 800k analyst ratings in the training set, and 200k analyst ratings in the validation set.

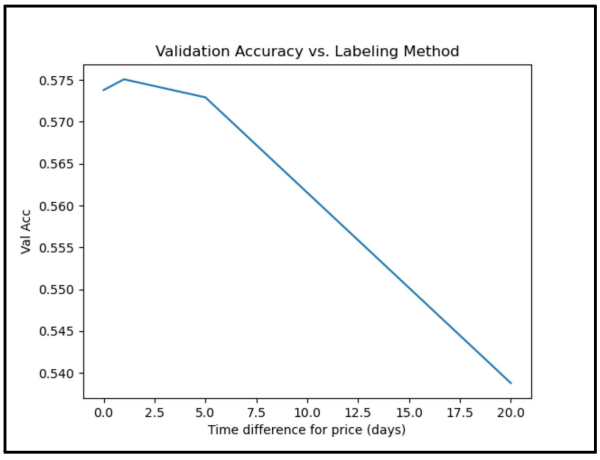

Figure 9: Effect of Time Difference on Stock Sentiment Labels

4.4.1 Implementation Details. For the implementation of this test, I use the same FinBERT model as from the previous test, with 0.2 dropout and a learning rate of $1e-5$. I trained a model for each labeling method using the Adam optimizer for 4 epochs and recorded the results.

4.4.2 Results. It is shown in figure 9 that predicting the next-day price difference for stocks using Analyst Ratings data had the highest accuracy out of all of the labeling methods. However, this only resulted in about a 0.1% increase in accuracy when compared to the dataset labeled with the same-day price difference. The next-five day price difference falls closely behind the same-day and next-day differences as well. Interestingly, the model had the lowest accuracy when predicting the next-twenty day price difference of the stocks, with an accuracy of 53%. Obviously, the information in the analyst ratings data is more relevant to the stock price in the time closest to when the ratings were published. Though for a long enough range, the price difference will usually tend to be positive, as the markets tend to increase over time with a 7-year average of 10% a year.

4.5 Ablation Studies

4.5.1 Testing the Dropout. Seeing as how the validation accuracy is unstable in most of these experiments, I wanted to experiment with various techniques for mitigating this. Obviously in all of my evaluations, I implement early stopping and select the model with the best validation accuracy, but I found this accuracy to vary largely between runs. Therefore, I test the dropout in the FinBERT model from 0.0 to 0.6 in increments of 0.1. This is done on the largest of the the three datasets I have examined, the analyst ratings dataset. I find that a dropout of 0.2 performs the best, though there is a difference of no more than .5% for all of the dropouts, as is shown in figure 10

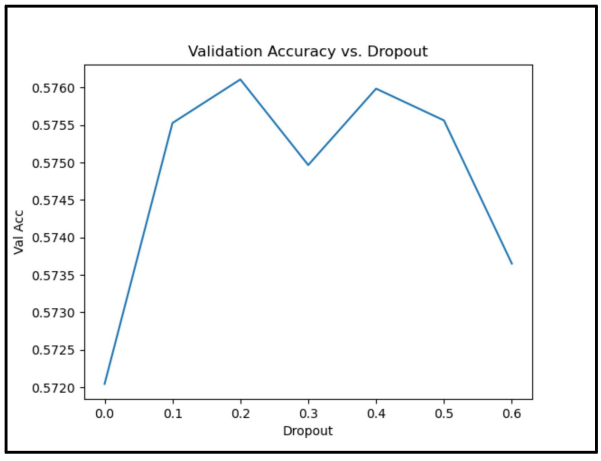

Figure 10: Dropout Test Results

4.5.2 Testing the Use of Stop Words. As a second ablation study, I examine if the removal of stop words (e.g. and, a, at, the) as well as numerical values. I conjecture that specific numbers may only add noise to the data, when all that is relevant is whether there is an increase or a decrease in the price.

Stop Words	Validation Accuracy
False	0.571
True	0.576

It is clear from the above table that the model performs better with the use of stop words and numerical values, though not by much. However, since the validation accuracy varies between runs, I cannot say for certain which approach is the best. Typically one would evaluate this with k-fold cross validation, however, I am dealing with a fixed train/validation split since I want to use information learned from past news headlines to predict the sentiments of future news headlines. It also may be a product of the fact that I am dealing with relatively short sequences, capping the maximum sequence length at 32 tokens.

4.5.3 Testing the Batch Size. I also tested the batch size, using a learning rate of $1e-5$ for 4 epochs. It is clear from figure 11 that the best batch size was actually 256, indicating the model may be able to generalize much better when using larger batches.

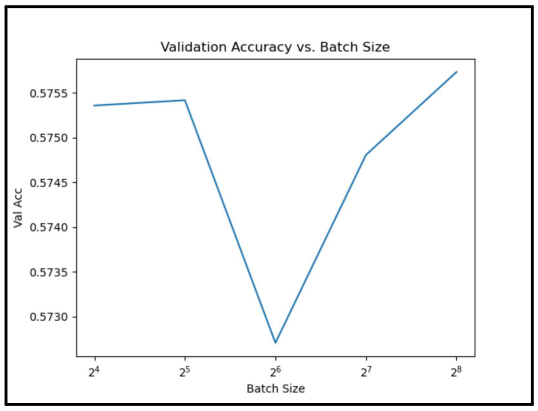

Figure 11: Batch Size Test Results

4.5.4 Testing the Learning Rate. Lastly, I tested the learning rate of the model, to see if the learning rate was causing the validation accuracy to not converge on any higher value. It is shown in figure 12 that the best learning rate was $1e-6$, however the learning rate of $1e-5$ which I was using for my main experiments was not much worse.

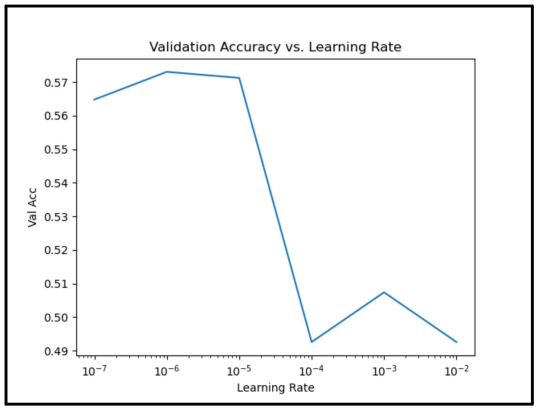

Figure 12: Learning Rate Test Results

5 DISCUSSION

It is clear that between all of the datasets and models tested for predicted the stock price direction, the predictions were not much better than random. However, those datasets which used exclusively financial data for individual stocks rather than generic world news data showed greater promise for predicting the price of stocks. It was also shown that many news headlines may not have much relevance for predicting the stock price, especially in the world news dataset. The predictions for the DOW Jones Industrial Average achieved 52.2% accuracy, while the predictions for stocks using Financial News achieved 52.8% accuracy, and the predictions for stocks using Analyst Ratings data achieved 57.4% accuracy. The fact that the pre-trained FinBERT sentiment models were able to achieve comparable accuracy indicates that not much information is being learned about the analyst ratings other than the sentiment of the ratings themselves. It is also clear that the LSTM model with GloVe embeddings is not well-suited for predicting the stock price direction, however this model may be more suitable for generic data such as world news data, rather than specialized data such as financial news. Models like this may be used when the size of the dataset is relatively small, and a massive amount of parameters aren't needed like in the FinBERT model (110M parameters). One important finding of this paper is that the short-term price differences in stocks may be more predictable using information like news data, rather than the long-term price differences. I think this is mostly due to the fact that analyst ratings and news about any given stock may lose relevance over time.

6 CONCLUSION

In conclusion, the findings of this paper are in support of the efficient market hypothesis that a stocks current price reflects all past information. However it must be examined how news data, such as that which was examined in these studies, can be combined with technical data and fundamental data to produce better results. Though, when using highly specialized data for a given stock, such as analyst ratings data, the results seemed to improve somewhat, achieving a maximum of 57.4% in accuracy. This information may be incorporated in the sentiment of the news headlines alone, and can be evaluated with similar accuracy using pre-trained FinBERT models with no extra finetuning, since these models were trained on financial-specific datasets. Lastly, these predictions may only be relevant in the short term, immediately after the articles were published. Therefore, in the future I think a much larger amount of data must be used, and a greater quality of data which includes other important indicators used by investors today.

REFERENCES

[1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *CoRR* abs/1908.10063 (2019). arXiv:1908.10063 http://arxiv.org/abs/1908.10063

[2] Girija V Attigeri, Manohara Pai M M, Radhika M Pai, and Aparna Nayak. 2015. Stock market prediction: A big data approach. In *TENCON 2015 - 2015 IEEE Region 10 Conference*. 1–5. https://doi.org/10.1109/TENCON.2015.7373006

[3] bot developer. 2020. Daily Financial News for 6000+ Stocks. https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests

[4] Eugene F. Fama. 1965. Random Walks in Stock-Market Prices.

[5] Eugene F. Fama. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance* 25, 2 (1970), 383–417. http://www.jstor.org/stable/2325486

[6] GennadiyR. 2020. Historical financial news archive. https://www.kaggle.com/datasets/gennadiyr/us-equities-news-data

[7] Ayman E Khedr, Nagwa Yaseen, et al. 2017. Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications* 9, 7 (2017), 22.

[8] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[9] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65 (2014).

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[11] Dev Shah, Haruna Isah, and Farhana Zulkernine. 2018. Predicting the Effects of News Sentiments on the Stock Market. In *2018 IEEE International Conference on Big Data (Big Data)*. 4705–4708. https://doi.org/10.1109/BigData.2018.8621884

[12] J. Sun. 2014. Daily News for Stock Market Prediction, Version 1. https://www.kaggle.com/aaron7sun/stocknews

[13] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *The IEEE International Conference on Computer Vision (ICCV)*.