Predicting Stock Prices Directly from Financial Articles

Ashwin Dubey
Georgia Institute of Technology
adubey38@gatech.edu

Pranav Manjunath
Georgia Institute of Technology
pmanjunath9@gatech.edu

Pankaj Dahiya
Georgia Institute of Technology
pankaj@gatech.edu

Kritika Venkatachalam
Georgia Institute of Technology
kvenkata8@gatech.edu

Abstract

Investment professionals and novices have long attempted to predict stock market returns from different factors. The techniques used for these prediction tasks range from fundamental analysis of equities to quantitative analysis of market factors. Natural language processing (NLP) has emerged as a quantitative field that allows machines to understand text and spoken words. Given the advancement of natural language processing techniques in the past 10 years or so, we attempt to apply various NLP based models to the stock market prediction task. Our results point to the emerging trend within finance of integrating complex quantitative language models into the investment decision making process.

1. Introduction

1.1. Background

Sentiment analysis on financial data is an increasingly popular topic in the fields of finance and data analysis. It refers to the use of NLP and Machine Learning (ML) techniques to analyze subjective views (opinions) in financial literature, news, and social media. Such analysis can help potential investors gain valuable insights into various market trends, and corroborates the fact that investors do make investment decisions based on finance-related news articles, especially negative articles [1].

Financial jargon is quite extensive (several online guides and books have been and are written just to help people get a handle on the exhaustive vocabulary that the world of finance offers) and often gives a different meaning to traditional English words, thus changing the sentiment of those words as well. One example is the word “share.” In traditional English, the act of sharing something with someone tends to indicate a positive relationship between those two individuals. However, in the world of finance, the word “share” refers to a type of asset, which in itself has no positive or negative sentiment and should therefore be identified as having neutral sentiment. Thus, NLP models trained on financial news articles and posts can usually only be used for financial analysis, and NLP models trained on other types of articles and posts can usually not be used for financial analysis. By extracting financial jargon from several 10-K filings, Loughran and McDonald were able to create a dictionary of negative, neutral, and positive finance-related words that can be used to speed up the process of gauging the tone of a particular finance article [2].

Previous literature regarding sentiment analysis on financial data has demonstrated that sentiment analysis can be used to forecast stock prices [3] and market trends [4]. Additionally, Tetlock et al. found that using sentiment analysis to find the fraction of negative words in a firm’s news stories can enable analysts to forecast the firm’s annual earnings [5].

1.2. Dataset Description

The dataset we used is a Kaggle dataset with stock price data scraped from Yahoo Finance articles and news data scraped from another Kaggle dataset. There are two csvs in the dataset, one corresponding to stock predictions and one corresponding to the actual stocks. We only used the stock prediction csv. In the stock prediction csv, the stock data is presented as a 13-column csv file. The first column, index, represents the relative recency of the entry in order from most recent to oldest. So the most recent entry is at index 0. The ticker column corresponds to the stock’s ticker symbol. The only stock that the csv contains data for is Apple, whose ticker symbol is AAPL. The date column contains the date formatted as yyyy-mm-dd. The category is the type of article: either news or opinion. The title column is essentially the headline while the content column is the description of the headline. The open column contains the

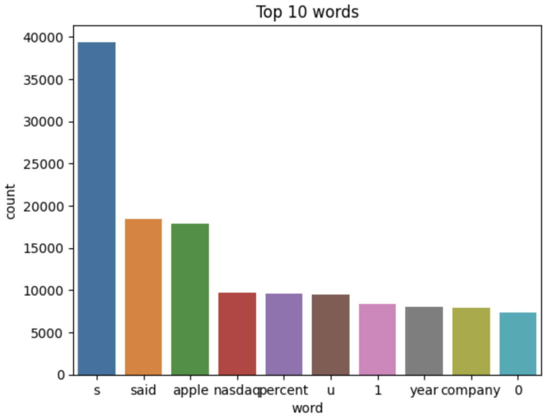

Figure 1: Top 10 Words in Dataset

opening price, while the high and low columns represent the highest and lowest prices respectively of the stock throughout the day. The close column contains the closing price of the stock for that day. The Adj. close column contains the closing price of the stock for the day after adjusting for paid-off dividends. The volume column contains the volume of the stock: the number of shares of that stock that were traded between the opening and closing of the stock that day. The label column contains two values: 1 if the close price of the stock is greater than its open price, else 0.

2. Model Architecture and Performance

2.1. Embeddings

GloVe contains non-contextual word embeddings while BERT contains contextual word embeddings. However, neither could be successfully utilized in our case because, as described in the Background section, financial jargon is too divergent from traditional English for us to use these embeddings. Word embeddings pre-trained on finance news articles could theoretically be utilized to improve our F-1 score, but owing to time constraints and the fact that we could only find one repository containing pre-trained word embeddings for financial jargon, we were not able to implement these. We did try using BERT just to see what our results would be like, but we only got an F1 accuracy of 50 %.

2.2. Logistic Regression, Random Forest, Support Vector Machine, Naive Bayes

Logistic Regression is a supervised-learning classification model for categorical variables. Random Forest Classification is a supervised ensemble learning technique that classifies based on the average result from several decision trees. Support Vector Machines are supervised learning algorithms that attempt to find hyperplanes in an N-dimensional space where N is equivalent to the total number of features in the data. Naive Bayes supervised-learning algorithms utilize Bayes' theorem and the conditional independence between every pair of features to make their classification decisions. These four models were the initial models we used to try and predict the stock labels.

2.3. Basic Model Performance

In our case, because we were trying to predict the label of the stock and there were only two possible options (0 or 1), all of the basic models we used were binary. We first ran all four models on both opinion and news articles. While all four models performed similarly, Logistic Regression performed the best with an F1 accuracy of 57.0 %, while Random Forest Regression came in a close second with an F1 accuracy of 56.7 %. SVM and Naive Bayes weren't too far behind at 56.1 % and 54.8 %, respectively. While we did expect SVM to be more accurate than Naive Bayes, we did not expect it to be less accurate than Random Forest Regression. This is because Wang et. al used several different binary ML classifiers to classify the sentiment of various StockTwits tweets and in their results, the SVM proved to be more accurate when compared to Decision Trees (which is what Random Forests are composed of) and Naive Bayes [6].

We then filtered out opinion articles because of our initial assumption that they might not contribute as well to market analysis as regular news articles, but our results weren't significantly different. Although the SVM did prove to be more accurate than Random Forest and Naive Bayes as we expected based on the experiments conducted by Wang et. al [6], the overall F1 accuracy results were pretty similar. The SVM's F1 accuracy was 57.3770 %, a slight improvement from 56.7 % when the opinion articles were included. The Random Forest model's performance also improved very slightly, producing an F1 accuracy of 57.1038 %. The Naive Bayes model's F1 accuracy was 56.9217 %, also a slight improvement from 56.1 %, while the Logistic Regression model's performance improved ever so slightly from 54.8 % to 55.6466 %.

2.4. Long Short-term Memory Network, Encoder-Decoder Model, Feedforward Neural Network

Feedforward neural networks are a simple type of neural network where information is only passed forward from input nodes, through hidden layers, and exits via output nodes. Long short-term memory (LSTM) networks are a special type of recurrent neural network (RNN) that are capable of combating the RNN's vanishing gradients issue by retaining information long-term. On top of having a hidden state at each timestamp, the LSTM contains cell states that contain long-term memory about the earlier parts of the se-

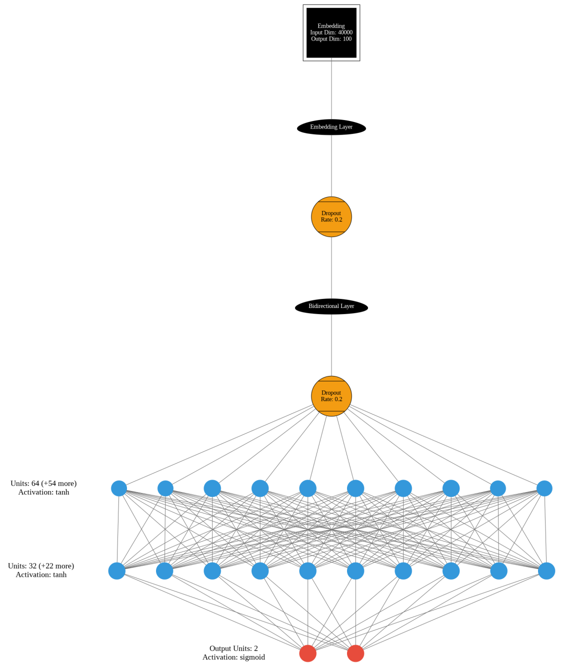

Figure 2: LSTM-FFNN Architecture

quence. In the Encoder-Decoder model, a sequence is embedded word-by-word and then passed into an encoder. The sequence is then encoded as a context vector and passed into a decoder, where the translation/response is formed based on a greedy or beam search algorithm.

2.5. LSTM, FFNN, Encoder-Decoder Model Performance

We set up our encoder as an LSTM and our decoder as an FFNN. The FFNN takes the encoder output and outputs 0 or 1 for the label. Our results were pretty poor as we plateaued at an F1 accuracy of 55 % after 5-10 epochs for both opinion and news articles. After filtering out opinion articles, our F1 accuracy actually got a little worse, plateauing at about 53 %. Note that the encoder model we utilized has performed well for traditional NLP tasks.

3. Existing Literature and Contextualization of Our Work

3.1. Existing Literature

Most of the existing literature on sentiment analysis in financial text focuses on modeling sentiments directly rather than predicting stock prices. The aim was to determine the sentiment of the text regarding a particular stock or the market as a whole, rather than forecasting the stock's price movement. Sentiment analysis models can provide insights into how the market may react to news or events and help traders make informed decisions. Thus, some recent studies have attempted to combine sentiment analysis with price prediction models.

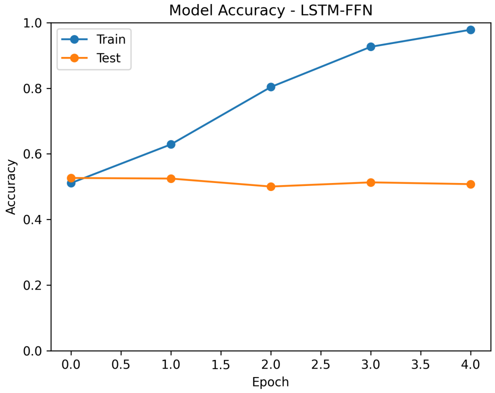

Figure 3: LSTM-FFNN Model Performance

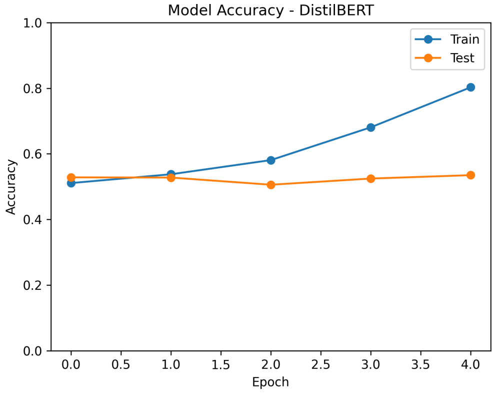

Figure 4: DistilBERT Model Performance

More recent studies have proceeded in a spirit of computer-aided objectivity which entails determining linguistic features to be used to automatically categorize text into positive or negative news. Davis et al. (2006) conducted a study to examine the impact of positive and negative language used in financial press releases on future firm performance. They found that readers tend to form expectations about the authors' biases and react more strongly to reports that contradict these expectations. Their findings suggest that readers, and therefore the markets, form expectations about not just the content but also the emotional tone of the text. Tetlock et al. (2007) focused on investigating the relationship between a pessimism factor, which was automatically generated from news text using term classification and principal components analysis, and its ability to

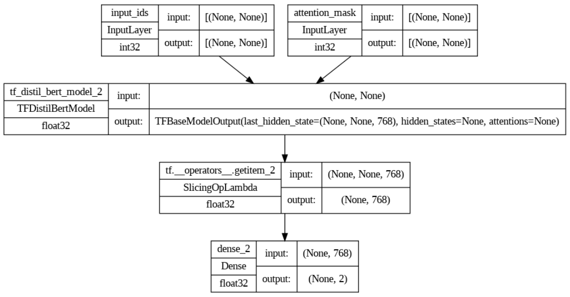

Figure 5: DistilBERT Model Summary (TensorFlow)

forecast market activity, specifically stock returns [5].

Overall, the existing literature on sentiment analysis in financial text highlights the potential of text-based sentiment analysis in forecasting market trends and informing trading decisions. However, further research is needed to develop more accurate and reliable models that can predict stock prices based on sentiment analysis.

3.2. ChatGPT

The popular new large language model ChatGPT has just started to be applied to market relevant tasks [9, 10]. Quantitative hedge funds have long used NLP to measure stock popularity from Twitter or news headlines [9]. Man AHL, a London based quantitative research fund, was first manually labeling sentences as positive or negative for different securities and using them for machine learning. Bloomberg states how Chat-GPT can pull off similar tasks without even being trained [9]. A recently published University of Florida paper gave ChatGPT financial news headlines and asked it to classify the headline as good, bad, or irrelevant to a stock price. The researchers then computed a numerical "ChatGPT score" and found a positive correlation between those scores and subsequent daily stock market returns [10]. They also state that more basic models such as GPT-1 , GPT-2 and BERT cannot accurately forecast returns, corroborating the results of our paper.

Models more sophisticated than ours cannot accurately forecast returns, which sheds light on the difficulty of the task at hand. In the context of our paper, we used entire documents to generate an "increase" or "decrease" label. It is entirely possible that the entire document is verbose and contains noise that could affect model performance. Note that our paper also generated price movement predictions directly from the data. In other words, the University of Florida paper classified a document's analysis and used that classification to compute a price movement. The approach to directly predict the movement could have resulted in poorer performance, but note that other, more sophisticated models like GPT-1 and GPT-2 also performed poorly as cited by Lopez-Lira and Tang [10].

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Naive Bayes	0.890	0.884	0.905	0.894
Random Forest	0.999	1.0	0.999	0.999
Logistic Reg.	0.999	0.999	0.999	0.999
SVM	0.831	0.955	0.776	0.856
LSTM-FFNN	0.989	0.990	0.989	0.989
DistilBERT	0.952	0.948	0.962	0.956

Table 1: Model performance metrics on Training Dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Naive Bayes	0.548	0.580	0.557	0.568
Random Forest	0.567	0.695	0.563	0.622
Logistic Reg.	0.570	0.592	0.579	0.587
SVM	0.561	0.807	0.549	0.653
LSTM-FFNN	0.507	0.532	0.536	0.534
DistilBERT	0.534	0.551	0.625	0.585

Table 2: Model performance metrics on Test Dataset

4. References

1 Arratia, A., Avalos, G., Cabaña, A., Duarte-López, A., Renedo-Mirambell, M. (2021). Sentiment Analysis of Financial News: Mechanics and Statistics. In: Consoli, S., Reforgiato Recupero, D., Saisana, M. (eds) Data Science for Economics and Finance. Springer, Cham. https://doi.org/10.1007/978-3-030-66891-4_9

2 Loughran, T., McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

3 L. Dodevska, V. Petreski, K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, and D. Trajanov, "Predicting companies stock price direction by using sentiment analysis of news articles," in Proc. 15th Annu. Int. Conf. Comput. Sci. Educ. Comput. Sci., Fulda, Germany, Jul. 2019, pp. 37–42.

4 C. Curme, H. E. Stanley, and I. Vodenska, "Coupled network approach to predictability of financial market returns and news sentiments," Int. J. Theor. Appl. Finance, vol. 18, no. 7, Nov. 2015, Art. no. 1550043.

5 Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S. (2008). More than words: Quantifying language to measure firm's fundamentals. The Journal of Finance, 63(3), 1437–1467.

6 G. Wang, T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, and B. Y. Zhao, "Crowds on wall street: Extracting value from collaborative investing

platforms," in Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput. CSCW, 2015, pp. 17–30.

7 https://github.com/sid321axn/bank_fin_embedding

8 Huosong Xia, Yitai Yang, Xiaoting Pan, Zuopeng Zhang, and Wuyue An. 2020. Sentiment analysis for online reviews using conditional random fields and support vector machines. Electronic Commerce Research 20, 2 (Jun 2020), 343–360. https://doi.org/10.1007/s10660-019-09354-7

9 Lee, Justina. "CHATGPT Can Decode Fed Statements, Predict Stock Moves from Headlines." Bloomberg.com, Bloomberg, 17 Apr. 2023, https://www.bloomberg.com/news/articles/2023-04-17/chatgpt-can-decode-fed-speak-predict-stock-moves-from-headlines?srnd=premium&sref=KkPzpZvz&leadSource=uverify+wall.

10 Lopez-Lira, Alejandro, and Yuehua Tang. "Can CHATGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." SSRN, 10 Apr. 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412788.

11 https://www.kaggle.com/datasets/deepakjoshi2k/yahoo-stock-prediction-by-news