CSCE 771: Computer Processing of Natural Languages
Prof. Biplav Srivastava, Fall 2020

Final Project Report: Using News to Predict Company Stock Movements

Hong Yung Yip (M13315513)

Background

Learning to predict market behavior based on historical data remains an open challenge despite the advent of Big Data and machine/ deep learning technologies [1]. Although there has been adoption of computer-aided stock trading [2], increasing research on designing profitable strategies [3-5], algorithmic modelling of the stock market, the *true state* as well as the *transition dynamics* of the stock environment remain unsolved due to the inherent complexity, adaptive, and temporal nature of the stock markets. Stock trading is a constant and iterative process of testing new ideas, tuning to public perceptions' towards major news events, receiving market trends (up/ down/ stay) as feedbacks, and from these experiences, improve and optimize the (machine/deep) learning algorithms over time [4]. This trial-and-error approach to decision making coupled with the *partially observable* factors such as news make this an interesting and well-suited problem for Natural Language Processing (NLP) with Deep Learning (DL).

Problem

Studies to accurately predict the (company) stock market remains an on-going research due to its volatile and non-linear nature. Despite existing studies on using supervised (eg. ANN, RNN, and LSTM [7]), unsupervised approaches (eg. K-means) [6], and reinforcement learning [2-4], results were still far from satisfactory mainly because relying on historical prices (open, high, low, and close) and trading indicators (buy and sell volume, market trends, and moving averages) alone were not enough to be an indicative of future market behavior. At the granular level, one of the main factors driving a company stock prices is investors' demand. Demand can be influenced by market's (investor and consumer) expectations and emotions. The voluminous and ubiquity of data (news data) today has enabled individuals at any scale to make analytical investment decisions. However, consuming, interpreting, and mining these data to determine useful signals can be challenging.

Therefore, the hypothesis of this project are to (1) determine whether we can leverage the news data (both title and description as separate inputs with respective encodings and embeddings) to predict a company stock movements (multi-class outputs: stay, up, or down), and (2) understand a particular stock sentiment with sentiment analysis [8] and how they can be used as discounting factors (cues) for stock movements based on principle that, if the news sentiment is positive, the stock will trend up, and vice versa, if the news sentiment is negative, then stock may trend down.

Approach & Solution

The main objective of this project is not to predict the stock prices (regression with numerical outputs), but rather, to learn and predict the stock trend movements (multi-class classification with three movement categories: stay, up, or down) at a holistic level. The contributions of this project are three-fold:

1. **Baseline:** Preliminary approach with supervised deep learning techniques such as Convolutional (CNN) and Long-short Term Memory (LSTM) networks [9] to determine whether news text (a combination of news title and news description) can be used as the natural language counterpart to predict a company stock movements
2. Determine the effects of news sentiment polarity (as additional features on top of the news text in (1)) as discounting factors in predicting a company stock movements
3. Evaluate the performance difference between the two unsupervised pre-trained language models: GloVe's (context-based) and BERT (transformer-based) embeddings in understanding news text

The scope of the project is further divided into four components: (a) news and financial data collection, (b) data preprocessing, (c) text representation and language models, and (d) system (neural network architecture) design. Figure 1 illustrates the end-to-end pipeline from data collection to preprocessing to model training and evaluation.

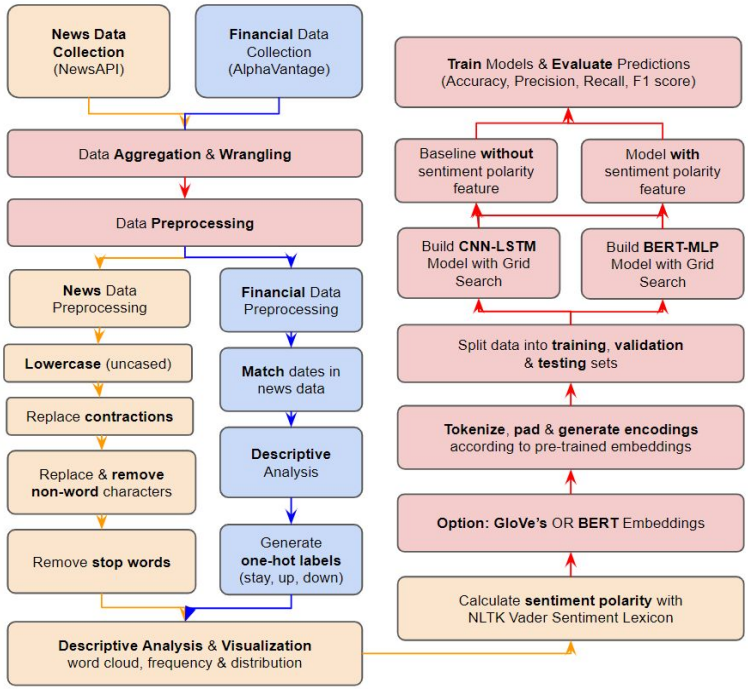

Figure 1: End-to-end news & financial data collection, preprocessing, model training and evaluation pipeline.

(a) News and Financial Data Collection

News and financial data for the period (10/9/20 - 11/9/20) are collected for the following stocks: (1) AAPL (Apple), (2) BA (Boeing), and (3) TSLA (Tesla) from NewsAPI (news) and AlphaVantage (daily price) respectively[1] (Table 1). The three stocks above experienced volatility in the past month due to major news events such as Apple released iPhone 12, Presidential elections, and COVID-19 which makes an interesting task for the deep models. The daily stock prices consist of Open, High, Low, Close, and Volume. The daily *open* price is used as the output label as it accounts for price movement based on the news for the entire (previous) day including after-market trading hours.

Table 1: Statistics of the news and financial data of each company respectively

News and Financial Data for the period (10/9/20 - 11/9/20)							
Sources: bloomberg, cnn, cnbc, business-insider, financial-post, fortune, hacker-news, time, wired, techcrunch, techradar, the-verge, the-washington-times, engadget, mashable, ars-technica, reuters, google-news, newsweek, next-big-future, politico							
Stock	**Keywords used to collect news data**	**Number of news collected**	**Size of vocabulary**	**Average news title length**	**Average news description length**	**Lowest (Open) Price**	**Highest (Open) Price**
AAPL (Apple)	apple, aapl, iphone, 5G	1499	7758	28	40	109.11	125.27
BA (Boeing)	boeing, airline, airlines, aviation	828	5347	26	36	145.75	179.0
TSLA (Tesla)	tesla, tsla, elon, musk	477	4028	27	39	394.0	454.44
Example News Data (Title and Description)							
News title: Apple just announced its 5th stock split in history[2].							
News description: Apple will split its stock for the fifth time in its history, the iPhone maker said on Thursday after reporting strong earnings that beat analysts' estimates. Apple's stock price will be quartered, from about $400 now to about $100 when the split happens.							

[1] The data are only collected up to a month period due to the limitation imposed by NewsAPI on developer account.
[2] https://markets.businessinsider.com/news/stocks/apple-stock-split-price-what-that-means-how-many-shares-2020-7-1029455711#

(b) Data Preprocessing

The collected news text data (both title and description) are subjected to a series of standard NLP preprocessing pipeline (Figure 1) prior to representation and modelling. Since the uncased version of the language models (GloVe and BERT) are used, the news text data are (1) lowercase, (2) contractions replaced, (3) non-word characters (symbols, punctuations, white spaces, and tabs) replaced, (4) and stop-words removed using a combination of both rule-based approach and the Natural Language Toolkit (NLTK) library. Then, a series of descriptive analysis such as word cloud, frequency, and distribution are performed to visualize the news content.

On the contrary, the daily open price (output label) is one-hot encoded into three classes (stay, up, or down trend) by calculating the difference between the open price today and previous day. Due to the small datasets with a one month period, we propose the following labeling schema to balance the (three) number of classes within these datasets: if the difference is greater than 2, it is labeled as uptrend; less than 2 as downtrend; and anything between is neutral (stay).

(b1) Sentiment Detection Algorithm

The NLTK Vader (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analysis[3] library is used for automatic sentiment detection of news titles and descriptions. It is a type of (pre-trained dictionary-based) sentiment analysis that is based on lexicons of sentiment-related words, which takes on two forms: (1) polarity-based, where pieces of texts are classified as either positive or negative, and (2) valence-based, where the intensity of the sentiment is taken into account. In other words, each word in the lexicon is rated as how positive or negative. It analyzes a piece of text by (tokenizing and) checking the number of words that are present in its lexicon and produces four sentiment metrics (positive, neutral, negative, and compound). The **compound** score (used in this project as the corresponding news title and description sentiment polarity) is the sum of all of the lexicon ratings which have been standardised to range between -1 and 1.

(c) Text Representation and Language Models

In this project, two language models are adopted and performance-compared: (a) **GloVe** (Global Vectors for Word Representation) [10] and (b) **BERT** (Pre-training of Deep Bidirectional Transformers for Language Understanding) [11]. Both models are architecturally different and are highly known for their great performance in the downstream NLP tasks. **GloVe** is a popular light-weight pre-trained word vectors (that is aggregated global word-word co-occurrence statistics) from large corpus such as Wikipedia [10], whereas **BERT** is one of the current state-of-the-art multi-layer bidirectional Transformer encoder with the “masked language model” (MLM) pre-training objective [11]. Due to the different representational requirements between the two language models, the news title and description are encoded respectively: word to integer dictionary mapping for **GloVe** and BERT tokenizer to generate the required input ids, attention masks, and segment ids for **BERT** model.

(d) System (Neural Networks Architecture) Design

Existing research on predicting numerical stock prices [1-2, 5-7] and sentiment analysis [8-9] yielded promising results, this project aimed to combine sentiment analysis (specifically polarity) with historical stock prices to predict a company stock trends (based on news data). Hence, the following experimental setup are performed to predict stock trends for AAPL, BA, and TSLA:

1. Baseline CNN-LSTM model (without sentiment polarity) with GloVe’s embeddings
2. CNN-LSTM model (with sentiment polarity) with GloVe’s embeddings
3. Baseline BERT model (without sentiment polarity) with fully-connected dense layers for fine-tuning
4. BERT model (with sentiment polarity) with fully-connected dense layers for fine-tuning

According to [9], a joint CNN and RNN (including GRU and LSTM variants) model (paired with Word2Vec and/or GloVe’s embeddings) produced the highest performance against standalone CNN as well as RNN models. The justifications are (1) CNN is able to retain the local features and sequential relations in a sentence, and when combined with RNN, (2) RNN can learn the long-term dependencies and the positional relation of features as well as

[3] https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

global features of the whole sentence. Whereas, BERT (this project uses the BERT-base with 12 layers and 768 hidden states) models (without and with sentiment polarity) are trained due to their recent triumphs in various NLP tasks [11]. Figure 2 below shows the end-to-end model training to testing pipeline.

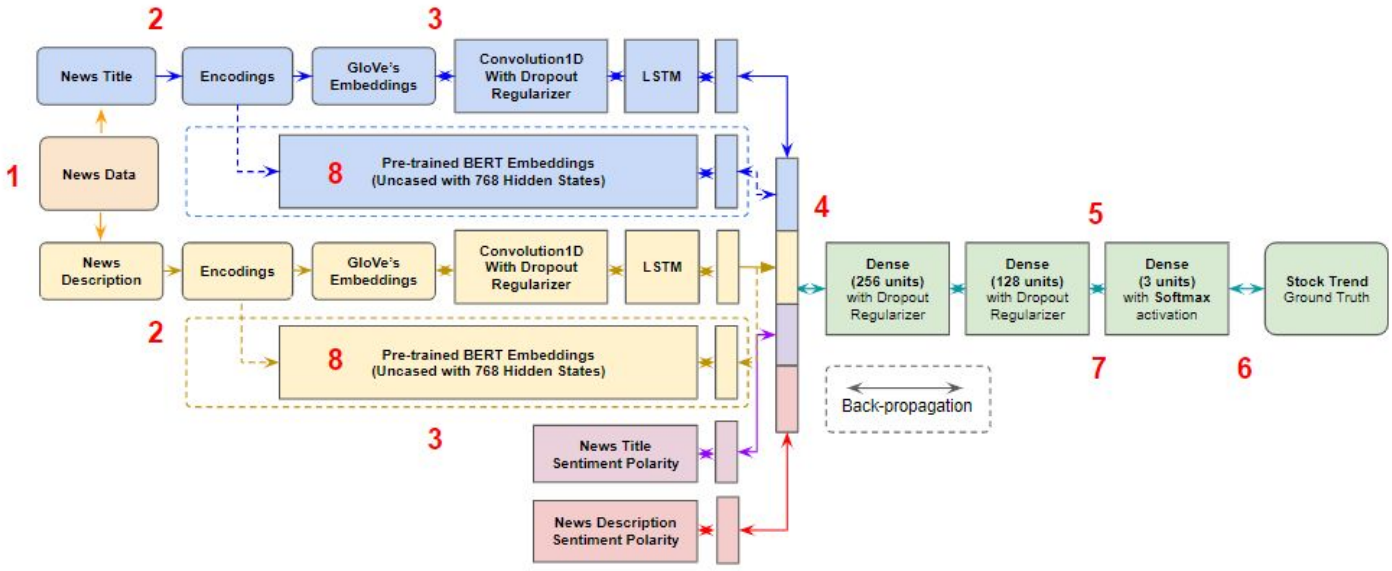

Figure 2: CNN-LSTM and BERT-MLP (Multi-layer Dense Perceptron) Architecture

1. News data are cleaned and preprocessed into news title and description respectively.
2. News title and description then are encoded into their respective GloVe' embeddings as separate inputs into the CNN-LSTM models.
3. The data are split into training, validation, and testing sets with the 80, 10, 10 rule respectively.
4. The hidden states from the CNN-LSTM models are concatenated with the news title and description sentiment polarities.
5. The concatenated vectors (matrices for batch) are subsequently fed into a series of fully-connected dense layers with variable hidden units and then lastly, an output layer with 3 output units representing softmax probabilities of [stay, up, down] trend.
6. The models are back-propagated with stochastic gradient descent and/or Adam optimizer with variable learning rates and with early stopping (when validation loss starts to diverge and explode).
7. The hyperparameters (dropout, learning rate, number of hidden units, and the size (deepness) of each layer) are fine-tuned with the Grid Search method.
8. The BERT variants of the pipeline repeat the step 1 to 7 with pre-trained BERT embeddings.

Demonstration

The collected news and financial data (AAPL, BA, and TSLA), the Python notebooks (Both CNN-LSTM and BERT-MLP variants), as well as instructions to reproduce the work are available at:
https://github.com/Joeyipp/predict-stock-trends-news.

Evaluation

The metrics to measure the models performance are the standard multi-class classification evaluation metrics: accuracy, precision, recall, and F1-score (Table 2).

Table 2: Evaluation results (based on best performing epochs)

Company	Model	Training	Validation	Testing			
		Accuracy	Accuracy	Accuracy	Precision	Recall	F1
AAPL	CNN-LSTM (Base)	0.3935	0.2118	0.3	0.3	0.3	0.3
	CNN-LSTM (Sentiment)	0.5262	0.5911	0.307	0.307	0.307	0.307
	BERT (Base)	0.4076	0.4138	0.32	0.32	0.32	0.32
	BERT (Sentiment)	**0.5725**	**0.6798**	**0.47**	**0.47**	**0.47**	**0.47**
BA	CNN-LSTM (Base)	0.3760	0.4186	0.337	0.337	0.337	0.337
	CNN-LSTM (Sentiment)	**0.4086**	**0.4286**	0.337	0.337	0.337	0.337
	BERT (Base)	0.3002	0.2411	0.357	0.357	0.357	0.357
	BERT (Sentiment)	0.3365	0.3875	**0.393**	**0.393**	**0.393**	**0.393**
TSLA	CNN-LSTM (Base)	0.4808	0.4769	0.479	0.479	0.479	0.479
	CNN-LSTM (Sentiment)	0.5165	**0.4923**	**0.521**	**0.521**	**0.521**	**0.521**
	BERT (Base)	0.6621	0.3846	0.458	0.458	0.458	0.458
	BERT (Sentiment)	**0.7912**	0.3846	0.458	0.458	0.458	0.458

Table 2 shows the evaluation results in terms of training, validation accuracy, and testing accuracy, precision, recall, and F1-score for three company stocks (AAPL, BA, and TSLA). In general, the models (both CNN-LSTM and BERT) with news sentiment polarity performed consistently better, with an average of 1.2-1.6 times better than news text without sentiment polarity.

Discussion

While the performance (trend) of the models are consistent across three different (AAPL, BA, TSLA) datasets from different industries, some of the observations include:

- Pre-trained BERT embeddings with fine-tuning performed on average, better than GloVe's embeddings.
- Models with news sentiment polarity perform on average 1.2-1.6 times higher than models without, which seems to imply and support the initial hypothesis that news sentiment (and intensity) does contribute to the fluctuation of stock movements.

Nonetheless, the results are far from state-of-the-art benchmarks and some of the plausible explanations are:

- Due to the limitation of publicly available data collection APIs, one month period of data collection is too short for any significant stock trends movement.
- The labeling schema (difference window) for trends movement is restricted due to the short datasets period.
- The random walk nature of the stock movements (in these datasets)
- The keywords used to search the news data play a crucial role in the quality of datasets being collected. As noted in the BA datasets, the keywords used (boeing, airline, airlines, aviation) are simply too broad. Upon sampling a subset of news data, there are many "noisy" news data such as advertisements, and paid articles. In contrast, the TSLA dataset which uses the keywords (tesla, tsla, elon, musk) are much more contextualized and specific, albeit smaller in size (477) compared to BA's (828). The results from the models of TSLA and BA differ on average, 1.5-2 times.

Some of the challenges include (1) quality news data collection, and (2) optimizing the models' hyperparameters. Due to the limited public APIs for news data, this project is only limited to one month period. Nonetheless, they are enough as preliminary datasets to test the experimental hypotheses. On the contrary, BERT embeddings require relatively high computational resources to train and fine-tune, hence optimizing the hyperparameters require a substantial amount of time. Nonetheless, the Grid Search method combined with the early stopping criteria (validation loss divergence) is used to test a range of different hyperparameter values.

Conclusion & Future Work

Accurately predicting future trends for a stock involves a number of factors and analysis such as fundamental analysis, technical analysis, and public sentiments. Nonetheless, predicting a company stock movements (trends) based on news data has its merits. This project aims to learn and predict the stock trends for three companies: Apple, Boeing, and Tesla with preliminary analysis and evaluation on one-month news and financial data, as standalone news data input and combining news data with sentiment polarity on two different language models (GloVe and BERT). Results illustrate the preliminary effects of incorporating news sentiment polarity with state-of-the-art language models such as BERT performed on average 1.2 to 1.6 times better than just text data, which reinforces the hypothesis that the market is (partly influenced) sentiment-driven. Due to the small datasets with short periods, **future work** may expand on (1) bigger historical datasets with longer periods, (2) keywords engineering (hand-crafting specific search terms during news data collection) to avoid "noisy" news data such as advertisements, and paid articles as well as (3) more attention on hand-feature engineering work such as data cleaning, filtering, scrutinizing on irrelevant and fake news, and data annotations. In addition, (4) more information can be tested and included in the models such as previous day's price change and to (5) improve on fine-tuning the models with better hyperparameters. Nonetheless, the results of this study provide a positive gleam on using news data for stock trends predictions.

Reference

[1] Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., & Fujita, H. (2020). Adaptive Stock Trading Strategies with Deep Reinforcement Learning Methods. Information Sciences.

[2] Carta, S., Corriga, A., Ferreira, A., Podda, A. S., & Recupero, D. R. (2020). A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning. Applied Intelligence, 1-17.

[3] Xiong, Z., Liu, X. Y., Zhong, S., Yang, H., & Walid, A. (2018). Practical deep reinforcement learning approach for stock trading. arXiv preprint arXiv:1811.07522.

[4] Yang, H., Liu, X. Y., Zhong, S., & Walid, A. (2020). Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. Available at SSRN.

[5] Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, 599-606.

[6] Powell, N., Foo, S. Y., & Weatherspoon, M. (2008, March). Supervised and unsupervised methods for stock trend forecasting. In 2008 40th Southeastern Symposium on System Theory (SSST) (pp. 203-205). IEEE.

[7] Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science, 170, 1168-1173.

[8] Bharathi, S., & Geetha, A. (2017). Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems, 10(3), 146-154.

[9] Wang, X., Jiang, W., & Luo, Z. (2016, December). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers (pp. 2428-2437).

[10] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.