Using News to Predict Company Stock Movements

Hong Yung Yip (M13315513)

Illustration of System in Action

https://github.com/Joeyipp/predict-stock-trends-news

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].

- Inherent complexity, adaptive, and temporal nature
- Constant and iterative process
- Public perceptions' towards major news events
- Market trends (up/ down/ stay) as feedbacks
- Optimize the learning algorithms over time

Problem

Results from existing studies [2-4, 6,7] were still far from satisfactory. Historical prices and trading indicators were not enough as investors' demand influenced by expectations and emotions. Big data enable analytical investment decisions. However, consuming and interpreting these data for useful signals can be challenging. The hypothesis of this project are to (a) determine whether we can leverage the news data to predict a company stock movements, (b) understand stock sentiment with sentiment analysis [8] and how they can be used as discounting factors.

Evaluation

1. The performance of the models (Table 1) are consistent across three different (AAPL, BA, TSLA) datasets from different industries.
2. Pre-trained BERT embeddings with fine-tuning perform on average better than GloVe's embeddings.
3. Models with news sentiment polarity perform on average 1.2-1.6 times higher than models without.

Discussion & Conclusion

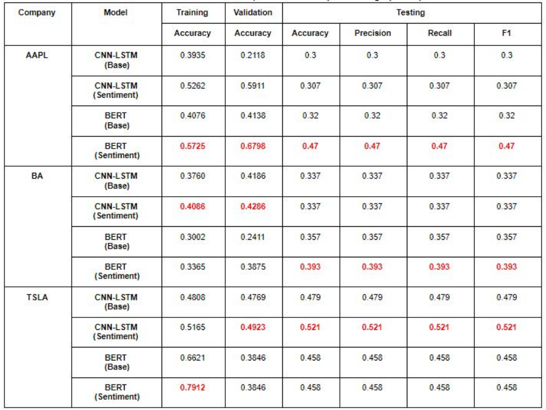
1. Limitation of publicly available data collection APIs.
2. Labeling schema for trends movement is restricted.
3. Random walk nature of the stock movements.
4. Search keywords affects the quality of datasets being collected.
5. Optimizing hyperparameters with Grid Search & early stopping.
6. Future work may expand on
    - Bigger historical datasets
    - Keywords engineering
    - Data scrutinization on irrelevant & fake news

Approach & Solution

The main objective is to learn and predict the stock trend movements (multi-class classification with three movement categories: stay, up, or down). The contributions are threefold:

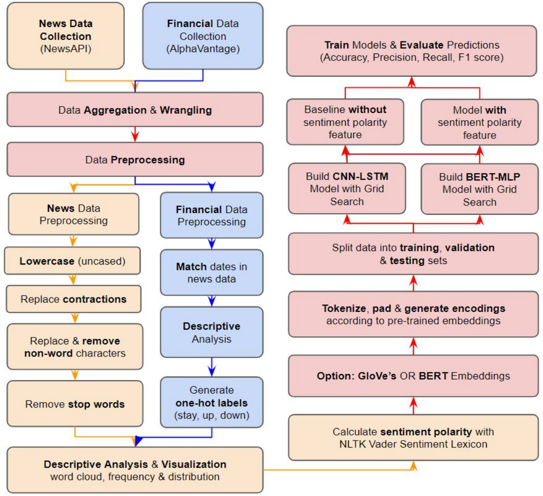
1. Baseline: Supervised approach with Convolutional (CNN) and Long-short Term Memory (LSTM) networks [9] to determine whether news text can be used as the natural language counterpart
2. Determine the effects of sentiment polarity (as additional features) to improve predictions
3. Evaluate the performance difference between the two unsupervised language models: GloVe's (context-based) [10] and BERT (transformer-based) embeddings [11] in understanding news text

The end-to-end approach (pipeline) for data collection, preprocessing, model training, and evaluation are illustrated by Figure 1 and 2.



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 1: Evaluation Results



News Data Collection (NewsAPI)

Financial Data Collection (AlphaVantage)

Data Aggregation & Wrangling

Data Preprocessing

News Data Preprocessing

Lowercase (uncased)

Replace contractions

Replace & remove non-word characters

Remove stop words

Financial Data Preprocessing

Match dates in news data

Descriptive Analysis

Generate one-hot labels (stay, up, down)

Descriptive Analysis & Visualization word cloud, frequency & distribution

Calculate sentiment polarity with NLTK Vader Sentiment Lexicon

Option: GloVe's OR BERT Embeddings

Tokenize, pad & generate encodings according to pre-trained embeddings

Split data into training, validation & testing sets

Build CNN-LSTM Model with Grid Search

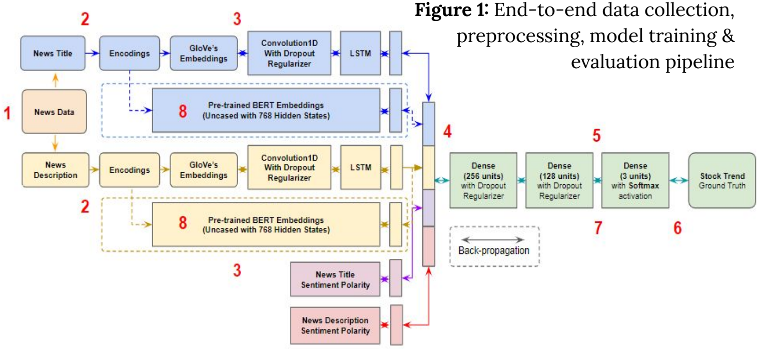
Build BERT-MLP Model with Grid Search

Baseline without sentiment polarity feature

Model with sentiment polarity feature

Train Models & Evaluate Predictions (Accuracy, Precision, Recall, F1 score)

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



2
3
News Title
Encodings
GloVe's Embeddings
Convolution1D With Dropout Regularizer
LSTM
1
News Data
8
Pre-trained BERT Embeddings (Uncased with 768 Hidden States)
4
5
News Description
Encodings
GloVe's Embeddings
Convolution1D With Dropout Regularizer
LSTM
Dense (256 units) with Dropout Regularizer
Dense (128 units) with Dropout Regularizer
Dense (3 units) with Softmax activation
Stock Trend Ground Truth
2
8
Pre-trained BERT Embeddings (Uncased with 768 Hidden States)
Back-propagation
7
6
3
News Title Sentiment Polarity
News Description Sentiment Polarity

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

Report & references available at
https://github.com/Joeyipp/predict-stock-trends-news/blob/master/reports/Final_Project_Report_YIP.pdf