Project Progress Report
Event-Driven Stock Prediction using NLP

Anant Jain, Ilya Yudkovich
CS6120 NLP Spring 2019
jain.anan@husky.neu.edu,
yudkovich.i@husky.neu.edu

Changes

Minor changes explained in detail in the Future Work section.

Preprocessing

We have created a large-scale financial news dataset by crawling Reuters.com. The dataset comprises of news for the past 1000 days for the top 37% market-cap companies (~2590 tickers) and is stored in a highly compressed pickle of a pandas dataframe. As expected, higher market-cap companies have more news as compared to lower market-cap companies, but since we are not concerned with the companies themselves but the type of events they undergo, we discarded ticker information from the data after aggregating the news by ticker and date (since a company can have multiple news articles in a single day).

Further, we normalized the text, eliminated stop words and filtered the dataset such that it is only left with words which appear in the vocabulary comprising the top 20000 most common financial words obtained from the Reuters corpus.

Next, in order to process the embeddings on a laptop, we limited the maximum sequence length per event to 1000 (something our laptops can handle) and padded the shorter sequences. This step ensured that we have a consistent 2D tensor for each event.

To build the labels for each event, we crawled historical stock prices (Open, Close, Adj Close, Date) for each ticker from Yahoo Finance and calculated their relative return w.r.t. S&P for that day. Further depending on the relative return value, we segregated the returns into categories (Strong Sell, Sell, Buy, Strong Buy) using quartile ranges and one-hot encoded them.

After merging the news and the labels datasets by date, we got our final dataset ready for some machine learning; comprising of a **reuters_sequences** column containing the 2D integer tensor of the filtered news of each event, and a **reuters_labels** column containing the one-hot encoded label indicating what would have been the best move to make given the news.

Method

Since we are using embeddings, we first prepared an embedding matrix for words in our financial vocabulary using the 42 billion version of pretrained glove embeddings each containing 300 dimensions.

Next, as we planned, we experimented with a density estimator as a baseline model, a 1D convolutional neural network to perform semantic composition over the input event sequence accompanied by a pooling layer in order to extract the most representative global features and associated them with stock trends through multiple shared hidden layers and an output layer.

All layers except the output use ReLU activation (the output uses Softmax activation). The model was compiled with categorical_crossentropy loss and a rmsprop gradient descent optimizer with accuracy metric.

Results and Evaluation

After letting the model train on 30637 samples and validate on 7749 samples, for a night; the results we got were very impressive for a baseline model. We achieved an average batch accuracy of 96% and a validation accuracy of 59.08% which is actually much higher than our expectations. We noticed when tested for a longer time, the performance gets worse. However, assuming no trading cost or liquidity issues, one technically can't loose money betting on a favourable game as long as the predictions are over 50% accurate.

The confusion matrix values obtained for each category (Table 1) also makes sense judging on how we segregated our labels. Moreover, it is evident from Table 1 that our classifier would have given way better accuracy, had we segregated our labels into just two categories: buy and sell.

Predicted Actual	Strong Sell	Sell	Buy	Strong Buy
Strong Sell	816	736	260	163
Sell	72	1380	327	151
Buy	57	408	1189	264
Strong Buy	40	271	422	1193

Table 1: Confusion Matrix for CNN 1D baseline model

Examples:

Ticker: AAPL.O

News: '''A U.S. federal judge has issued a preliminary ruling that Qualcomm Inc owes Apple Inc nearly $1 billion in patent royalty rebate payments, though the decision is unlikely to result in Qualcomm writing a check to Apple because of other developments in the dispute.'''

Prediction: "Buy"

Critique: The above news clipping instantly puts AAPL in good light even though it's unlikely to profit from the dispute. Nevertheless, it won't be a bad decision to Buy.

Ticker: AMZN.O

News: '''PARIS Casino's upmarket Monoprix supermarket chain is working to expand its partnership with E-commerce giant Amazon in France, following a successful launch in Paris, Monoprix's Chief Executive said on Thursday.'''

Prediction: "Strong Buy"

Critique: Definitely a strong buy. This is a no brainer.

Ticker: MSFT.O

News: SAN FRANCISCO Some Microsoft Corp employees on Friday demanded that the company cancel a $480 million hardware contract to supply the U.S. Army, with 94 workers signing a petition calling on the company to stop developing "any and all weapons technologies.'''

Prediction: "Sell"

Critique: Above news definitely puts Microsoft in legit pressure. Selling is the right action to take here.

What is working and What is wrong

For the data we have now, a CNN 1D with global max pooling isn't misbehaving too much and is surprisingly giving a solid performance if we just think about profit and loss.

However, as we know, when using embeddings, density estimators (convolution neural network) can achieve good results but can misbehave in high dimensions. Our current model is not something we can wholeheartedly rely on. Decreasing validation accuracy is a good sign indicating this. We need something better which can capture real-world market dynamics.

Future work

This is a very rough work. We can construct a better label by calculating relative return by comparing the stock price of the company and the corresponding industry, instead of comparing everything with S&P 500. It is almost like hedging, as long as an investor knows which company does well in some specific industry, he can make a decent prediction.

Also, we observed that some words have a strong effect (positive or negative) in finance e.g. merger, company acquisition, etc. The same was stated by Tim Loughran and Bill McDonald in 2009. Therefore, we plan to incorporate some degree of sentiment analysis to our model.

In addition to that, we plan to try a novel neural tensor network (NTN), which is capable of learning semantic compositionality over the arguments of an event by combining them multiplicatively instead of only implicitly, as with standard neural networks. We believe an NTN would be much more reliable than a density estimator like CNN for this task.

Also, as advised by the teaching staff, we plan to evaluate our results based on a Matthews correlation coefficient (MCC) besides the conventional measures like accuracy and precision.