Neural based event-driven stock rally prediction using SEC filings and Twitter data

Magdy Saleh, Surag Nair
{mksaleh, surag}@stanford.edu

Problem

- Gaining insight into the driver of stock price is a well-established difficult problem
- An approach outlined in Lee et. al, proves how adding text based features allows for some predictive power using random forests

Task Descriptions

Task 1:

- Given an earnings filing (8K) with SEC, we build models to predict the price movement in the next 24 hours

Task 2:

- Expand on the neural methods in task 1 and build a dataset with twitter information to study the impact of adding.
- Given any 8K filing for a company we leverage twitter data and the filing text to predict price movements over a 24 hour window

Data

Table 1. Features used across models

Earnings Surprise	The gap between consensus and reported earnings per share (EPS)
Trade Vol.	The number of trades for given ticker made during the day.
Volatility index	The volatility index value of the S&P 500 (ticker: VIX) at the market close before the 8-K report is released.
Tweets	All tweets on the day of the 8-K filing. Character based embeddings are learned using convolutional operators.
8-K	Multiple embeddings of 8-K text including GloVe based and learned embeddings.

- From Lee et al.[1] we obtain the initial dataset for task 1. It is composed of a total of **37k** examples that span the years **2002-2012**
- We construct a **new dataset** task 2. It spans the years of **2013-2018** and contains **~60k** examples with twitter data for each example

References

1. Heeyoung Lee, Mihai Surdeanu, Bill Maccartney, and Dan Jurafsky. On the Importance of Text Analysis for Stock Price Prediction. Technical report.
2. Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. 2015.

Models

For each task we try multiple different models. The best ones used are:

- **Task 1:** LSTM + GloVe[2] for 8-K text, a 2-layer FCC layer for non-text feats
- **Task 2:** CharCnn + Attn for tweets, LSTM for 8-K, 2-layer FCC for non-text feats

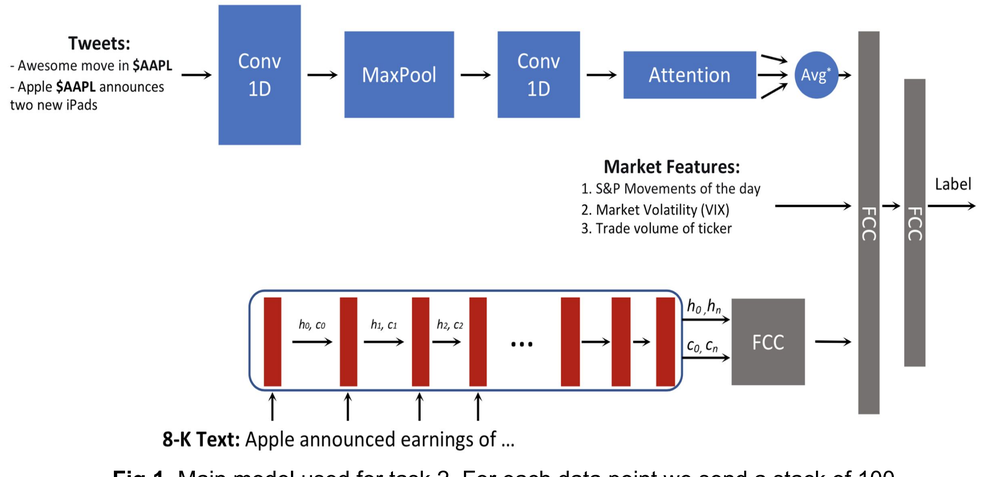

Fig 1. Main model used for task 2. For each data point we send a stack of 100 tweets through the CNN and then we compute an attention distribution over all the tweets passed in.

Attention on Tweets

- We rank our tweets by the attention scores provided by the model.
- In **Fig. 2** the top ranked tweets contain earnings results for the stock, which is a strong predictor of the price movement.
- The bottom ranked tweets are more general tweets that are bot generated and not specific to the stock in question.

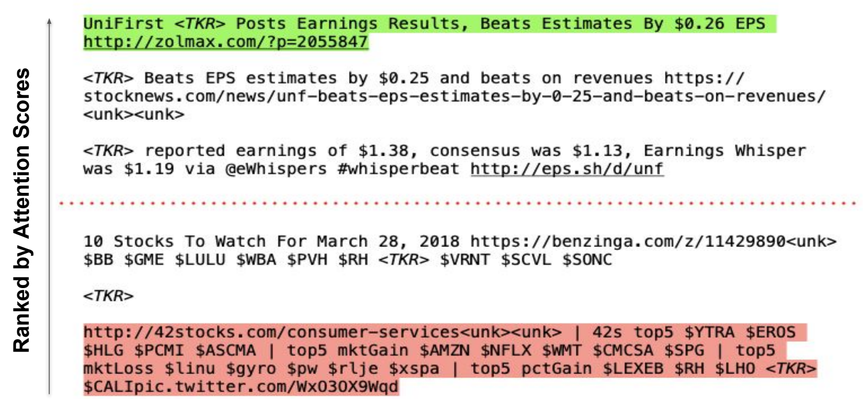

Fig 2. Tweets as ranked by the attention scores assigned by the model.

Results

We run a collection of models for both task 1 and task 2, **successfully beating our baselines** in both cases.

Table 2. Summary of results

Features	Model	Val. Acc.
-	Random Guess	33.33%
-	Majority Classifier	41.78%
VIX + Earnings	2-layer Feed Forward	49.29%
VIX + Earnings + 8-K	Bag of words for text + 2-layer FCC	47.23%
VIX + Earnings + 8-K	LSTM + 2-layer Feed Forward	51.30%
VIX + Earnings + 8-K	LSTM + GloVe + 2-layer FCC	**51.45%**
-	Majority Classifier	49.13%
VIX + Vol.	2-layer Feed Forward Network	49.56%
Tweets + 8K + VIX + Vol.	LSTM + CharCNN + 2-layer FCC	50.92%
Tweets + 8K + VIX + Vol.	LSTM + CharCNN with Attn. + 2-layer FCC	**52.61%**

Error Analysis

"SNX Beats EPS estimates by $0.11 and beats on revenues"

- Model predicts the price will increase on the back of this but the outcome was a price dip. High randomness

"Jefferies Group Raises American Eagle Earnings Estimates to $1.17 EPS (Previously $1.16).",

- Model fails to predict the price increase given the positive news. Language used is very specific to finance. Model needs more training on finance specific terminology,

Conclusion

- Hard problem, given high degree of randomness in the data
- New dataset scraped from multiple sources
- Clear advantage to the addition of text based features to the performance of both models
- Attention based tweet model learns to correctly prioritise information rich tweets that are relevant to stock price movements
- The ability to rank tweets and filter those that are below a cutoff threshold can potentially be a useful tool for investors.