Event-Driven Asset Pricing Prediction

Daniel Huang (dhuang7@stanford.edu), Stanford University Department of Economics

Abstract

The ability to predict future asset pricing will directly disprove the Efficient Market Hypothesis, a central tenet to the operations of financial markets. I attempted to use newspaper headlines to predict directional movement in the Dow Jones Industrial Average (DJIA), which tracks the performance of large American companies. I used both traditional natural language processing techniques with machine learning as well as various LSTM approaches, included stacked LSTM, for this classification problem. The LSTM model outperformed all other techniques, although further work is needed to generalize this result to all asset classes.

Predicting

We want to predict the price movement of the Dow Jones Industrial Average (DJIA), which measures the performance of large American companies, using news headlines as the input. If we can achieve a high prediction accuracy, this will disprove the Efficient Market Hypothesis, a central tenet of the operations of financial markets that would not allow stock price prediction to be possible [1]. The inputs are the newspaper headlines, from Reddit News, Reuters, BBC, and Yahoo Finance, from 2008 to 2016. The output is the direction of price movement of the DJIA, so this is a classification problem.

Data and Features

The time series data is discretized by day, and has been corrected for class imbalance issues. The features are news headlines from a variety of leading financial news sources, and have been scraped for a period of eight years (a summary of the dataset structure is found below). We use a tokenizer based on word frequency to convert textual data into workable vectors.

Dataset format:

Date	Label: DJIA price movement	Features: news headlines			
2008-06-08	1	Headline 1	Headline 2	…	Headline 50
2008-06-09	0	Headline 1	Headline 2	…	Headline 50
2008-06-10	0	Headline 1	Headline 2	…	Headline 50
…	…	…	…	…	…

One example:

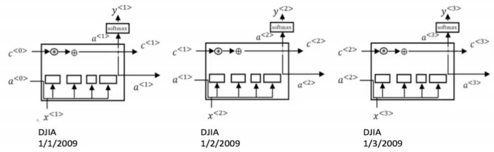
2013-11-09	1	German support for small business has kept its economy thriving as the rest of Europe languishes in recession	Once touted as an economic miracle, faltering India cannot provide jobs for millions of university graduates.	…

Baseline Model

For our baseline model, we used traditional NLP along with machine learning classification algorithms, including logistic regression, random forest classifier, SVC, linear SVC, and KNN.

LSTM Model

We tried implementing a feed-forward sequence model in the form of a LSTM. We used GridSearch for hyperparameter tuning, and tested different values for vocabulary size, activation function, loss function, optimizer, and number of epochs. In an LSTM model, each day is represented as one cell, with the day's collection of headlines and DJIA price movement.



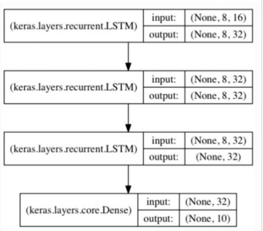
DJIA
1/1/2009

DJIA
1/2/2009

DJIA
1/3/2009

Stacked LSTM Model

We tried a stacking three LSTM models in order to better capture high-level temporal representations [2]. The first LSTM returns the full output sequence and feeds it to the second LSTM, and so on. The input sequence is converted into a single vector, and the temporal dimension is dropped from the output [3].



Results

We found that, for the baseline model, KNN performed the best with a 0.55 accuracy score. For the deep learning model, the single LSTM performed the best with a 0.58 accuracy score.

Component	Model	Accuracy Score
Baseline	Logistic Regression	0.45
Baseline	Random Forest Clf	0.52
Baseline	SVC	0.50
Baseline	Linear SVC	0.54
Baseline	**KNN**	**0.55**
Deep Learning	**LSTM**	**0.58**
Deep Learning	Stacked LSTM	0.56

Discussion

As expected, most of the models performed with an accuracy score near 0.50, which would suggest that they are as good as a random coin flip, which is what the Efficient Market Hypothesis would suggest. The best-performing model was the single LSTM model, which is not surprising since it can better learn complex textual relationships than can simple machine learning classification algorithms. We will have to investigate whether or not the 0.58 accuracy score is statistically significant enough to conclude that our model has economically predictive power. Otherwise, as expected, the Efficient Market Hypothesis holds.

Future

In the future, I would like to expand this problem to predict price movements of other assets, such as bonds and derivatives, as some asset classes are more prone to news headlines (ex: short-term maturity securities) than others, so it would be interesting to see if artificial intelligence can back up that intuition. Another interesting extension would be to convert this problem into a regression problem, to predict the actual price of the asset instead of only predicting the direction of price movement.

References

1. Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. Journal of economic perspectives, 18(3), 25-46.
2. Goldberg, Yoav. "A Primer on Neural Network Models for Natural Language Processing." J. Artif. Intell. Res.(JAIR) 57 (2016): 345-420. https://scholar.google.com/scholar?cluster=3704132192758179278&hl=en&as_sdt=0,5 ;
3. Keras. (n.d.). Getting started with the Keras Sequential model. Retrieved from https://keras.io/getting-started/sequential- model-guide/