Stock Price Prediction from News Headline Embeddings

Ryan Almodovar
Stanford University: CS230

Predicting

The goal is to predict the S&P 500 index based on news headlines published and the day. The inputs are the collection of headlines, transformed into word/sentence embeddings and processed to fit the delta change of the opening and closing price of the day. The models resulted in predicting in a 55%-62% accurate range.

Data

The data came from the UCI News Aggregator dataset from Kaggle, and the historical stock price data was downloaded from Yahoo Finance. The {0,1} label needed to be computed for each day by computing the delta of the opening and closing price of each day

Features

The features of the input are the word/sentence embeddings by applying a transformer on the headlines. Universal Sentence Encoder produces 512-dimensional vectors, word2vec uses 300-dimensional vectors

Models

These embedding is produced via different means from word2vec + CNN, BERT, and the Universal Sentence Encoder

BERT: MLM + softmax(Wx)
Word2vec + CNN:

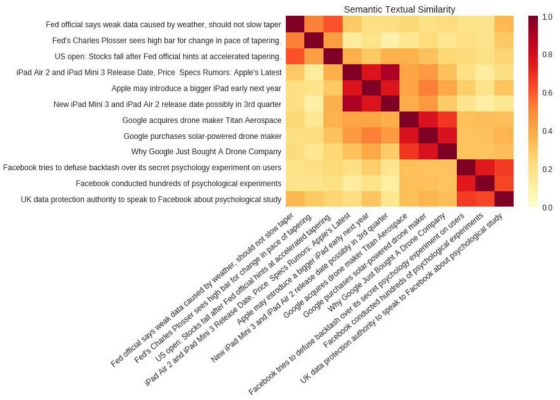
$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b).$$

Universal Sentence Encoder:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \left(1 - \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\,||\mathbf{v}||}\right)/\pi\right)$$

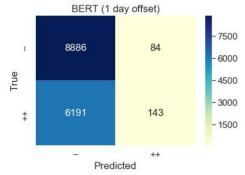
Results

Model	Train Error 146,731 total	Test Error 18,343 total
Word2Vec	35.2%	48.1%
Word2Vec + CNN	28.9%	37.8%
BERT	27.3%	39.8%
BERT (day offset 1)	25.5%	38.7%
BERT (day offset 2)	26.2%	41.2%



Discussion

The confusion matrices listed for each of the BERT day offsets reveals that each fine-tuned model on the training data tends to predict the market will decrease much more often than increase.
The matrix for the BERT-Base however tends to over predict that market will increase.
This can be due to the thousands of negative headlines being combined for each single day, and with the current implementation, each headline may have an equal effect of influencing the stock market rather than being properly weighted in comparison with more influential entities.



Future

For future work, if I had more time I would also want to augment BERT and Universal Sentence Encoder with more advance models on the embedding input vectors.
I would also look into factoring in other features rather than only the label, such as the delta values, time series analysis, or the categorized hostnames and draw further insights by incorporating those features in the overall model.

Youtube link:
https://www.youtube.com/watch?v=FXtP0cfgefY

References

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

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.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Sung, Y. H. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2823-2824). IEEE.