Using news titles to predict intraday DJIA movements

Arjun Arora, Avoy Datta, Victoria Ding

CS 224N | Winter 2018 - 19

CS 224N

Introduction

- We experimented with RCNN-based models to predict the intraday directional movements of the Dow Jones Industrial Average (DJIA).
- Prior work has shown that hybrid RNN and CNN architectures are effective at both capturing semantic information from financial news titles and modeling temporal dependencies.
- When it comes to linguistic input, most existing methods draw exclusively from financial news articles.
- We have attempted to extend this framework to *general* news headlines and experiment with attention mechanisms to better model temporal variations.

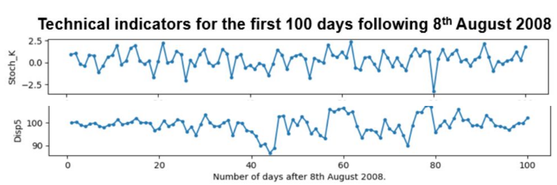
Dataset

- ~2000 days of data, from August 8, 2008 to January 7, 2016
- Financial information (high, low, close), 25 Reddit news headlines, movement label (increase/stay the same vs. decrease)) per day

Dataset	Positive label (1)	Negative label (0)
Entire dataset	0.534	0.466
Training set	0.542	0.457
Validation set	0.482	0.517
Test set	0.530	0.469

Features

- 7 technical indicators[2]
  Stochastic %K, Stochastic %D, Momentum, Rate of Change, William's %R, A/D, Disparity 5



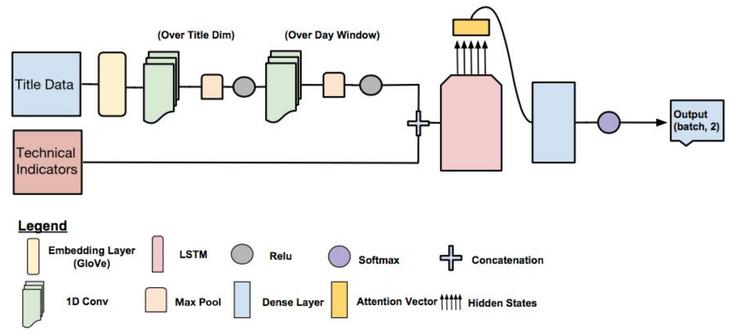
- 25 news headlines, encoded with 300-dim GloVe embeddings
- Then looked at both feature types (technical indicators and news headline) over the N previous days (N = 5)
  - Example titles, stopwords removed:
    [<start>, "Apple", "Under", "Fire", "Taiwan", <end>]
    [<start>, "Bush", "puts", "foot", "Georgian", "conflict", <end>]

Methodology

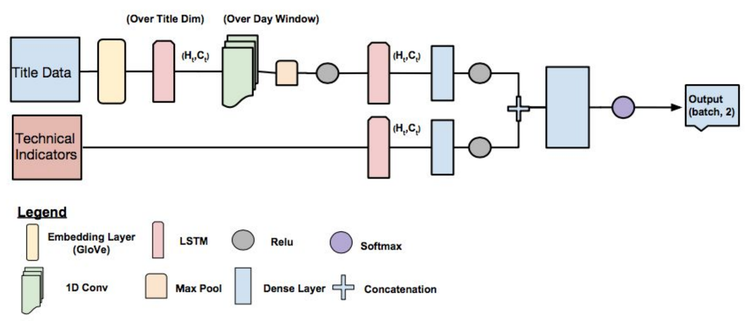
- Trained all RCNN models for 20 epochs
  - Train: 1600 samples, Validation: 200, Test: 180
  - Trained using LogSoftmax + NLL Loss
- SVM (radial basis) trained until convergence on each window
  - Each window (N=25,50) split into Train, Val,Test, ratio 8:1:1
  - Overfit results discarded after each window is finished
  - SVM accuracy averaged across all windows
- Both model families (SVM, RCNN) evaluated using binary classification accuracy

Experiments

RCNN with conv title embeddings and multiplicative attention:

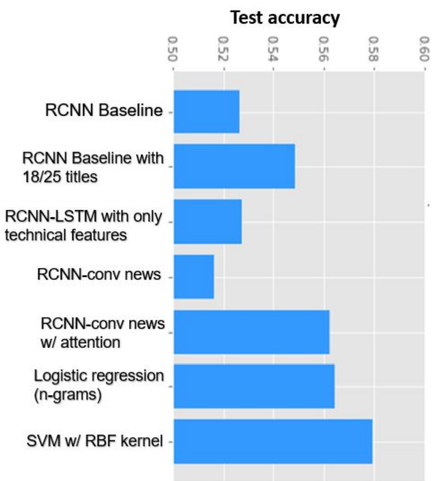


RCNN-LSTM with LSTM-based news headline embeddings:

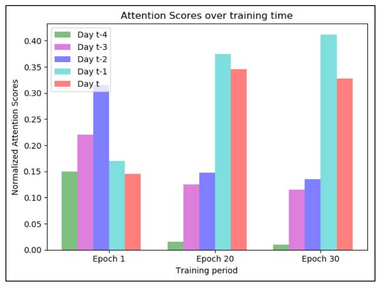


Results

Test accuracies of experiments



Attention scores for news title embeddings over 5 days



Analysis

- Experiments show that we beat baseline on our Reddit dataset (56 %)
  - Compared to Vargas et al, still shy ~6% accuracy
  - May be due to the inherent expressivity of S&P 500 vs. DJIA
  - Vargas used primarily financial news vs. general news
- Best model was the SVM
  - RCNN still cannot accurately capture sentiment
    - Cannot extract current DJIA movement from previous trends
  - It is better to just overfit to recent movement data
- Overall, financial time series data is hard to fit
  - SOTA is still only around 62.3% *on financial news only*
- Moving forward, if we were to try again, using BERT embeddings may be better than GLoVe, CNN, or LSTM embeddings
- Perhaps models more focused on financial statistics would do better than NLP based models

References

\

[1] M. R. Vargas, B. S. L. P. de Lima & A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, 2017, pp. 60-65. doi: 10.1109/CIVEMSA.2017.7995302
[2] Zhai, Y.Z., Hsu, A.L., & Halgamuge, S.K. (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. *ISNN*.