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.

Models

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

BERT: MLM + softmax(Wx)

\[ c_0 = f(W \cdot x_{i+k-1} + b). \]

Word2Vec + CNN:

\[ \text{sim}(u, v) = \left(1 - \frac{\|u - v\|}{\|u\| + \|v\|}\right).\]

Universal Sentence Encoder:

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Error</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>140.7, 313 total</td>
<td>18.3, 4 total</td>
<td></td>
</tr>
<tr>
<td>Word2Vec</td>
<td>35.2%</td>
<td>48.1%</td>
</tr>
<tr>
<td>Word2Vec + CNN</td>
<td>37.8%</td>
<td>38.7%</td>
</tr>
<tr>
<td>BERT</td>
<td>27.3%</td>
<td>39.8%</td>
</tr>
<tr>
<td>BERT (day offset 1)</td>
<td>25.5%</td>
<td>38.7%</td>
</tr>
<tr>
<td>BERT (day offset 2)</td>
<td>26.2%</td>
<td>41.2%</td>
</tr>
</tbody>
</table>

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 Baseline however tends to over predict that market will increase. This can be due to the thousands of negatives 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=FXiP0cYspiY

References


