Proposal: Relevance Adjusted Sentiment for Market News

Shrivu Shankar[1]

I. INTRODUCTION

This paper proposes the creation of a sentiment metric that takes in to account a document's relevance with respect to a target. More specifically the sentiment of a financial news article from the perspective of an arbitrary company.

Relevance aware sentiment aims to build on existing attempts for sentiment-based market prediction with the following advantages.

More data: More articles can be used to identify a company's sentiment without the downsides of irrelevant sentiment biasing its score.

Better Embeddings: Both articles and companies can be embedded using transformer-based architectures [1] that outperform bag-of-words based models. In addition, embeddings can be generated for companies.

Interpretability: By predicting relevant sentiment rather than whether a stock will go up or down, the model can be more intuitively acted on by a human investor and be used in human-in-the-loop trading strategies.

II. Related Work

Significant work has been done in the field of NLP and stock market prediction. A Google search (as of March 2020) for "predicting stock prices based on news articles" yields 163 million results and 1,728 repositories on GitHub. Many of these projects rely on bag-of-words style methods for sentiment [2], [3] [4] [5] and those that use more sophisticated encodings tend to lack relevant sentiment (sentiment calculated with respect to a market entity) [6] [7] [8] [9].

Ming et al. [10], uses sparse matrix factorization to find stock correlations between companies and articles. This method accounts for relevance but doesn't include sentiment.

Hu et al. [11] designed a custom deep learning model named "Hybrid Attention Networks" which uses "News-level Attention" to identify relevance. However, their method only provides relevance for a fixed target (i.e. their model can only be trained for one sector or company at a time).

III. Data

The dataset to be used was collected via webscraping [12] popular financial news websites and is split into two parts: articles with company labels (5%) and just articles (95%). The unlabeled articles will mainly be used for testing.

The core of the model will take in an article A_i and a company C_j and produce a relevant sentiment score s^* 1 where $sign(s^*)$ indicates positive or negative sentiment and $|s^*|$ corresponds to the relevance. Whether A_i is the text of the article, the headline, or both will be determined in the next section.

IV. Methodology: Article Embeddings

To produce the embedding of an article (e_{A_i}), three methods will be tested: Pretrained BERT [1], Doc2Vec [13], and bag-of-words (baseline). Each of these will be trained with only headlines, only content, and both headlines and content.

The validity of each method will be seen by visualizing the dimensionally reduced embeddings and inspecting patterns with known labels (e.g. articles about similar sectors should be grouped together) and by evaluating the accuracy of the task in the next section.

V. Methodology: Company Embeddings

The embedding of a company (e_{C_j}) 1 will be found by training a model to maximize the accuracy of predicting r ($r = 1$ if article i has the label company j else 0) s.t. $CosineSimilarity(e_{A_i}, e_{C_j}) = r$. This model may employ the use of additional fully connected layers attached to e_{A_i} if it increases the accuracy of the model by adjusting the precomputed article embeddings.

The binary accuracy of predicting r (baseline 50%) from known labels can be used to evaluate these embeddings.

VI. Methodology: Sentiment

Sentiment (s) will be tested using VADER [14], Pattern [15], BERT pretrained on a sentiment tasks, and a unsupervised approach using historical price data s.t. articles are labeled positive if the companies they are related to did well (a stock price increase) and vise versa.

This will be evaluated by how well related articles' sentiment correlates with a company's price movements. 2

$$s^* = s \cdot CosineSimilarity(e_{A_i}, e_{C_j}) \tag{1}$$

VII. Preliminary Results

With 60k labeled articles, article and company embeddings were generated along with sentiment using only BERT and Pattern. The company embeddings show learned similarities between companies that operate in the same sector 1 and the binary accuracy of r was 85%. Using the larger dataset $s*$ was calculated for every article and company for that last few months. Plotting $s*$ overtime yielded expected results 2 in that Netflix (online video streaming) has a significantly higher relevant sentiment than Disney (theme parks) in the onset of the COVID-19 pandemic outside of China.

[1] Shrivu is an undergraduate in Computer Science at The University of Texas at Austin. Email: shrivu.shankar@utexas.edu

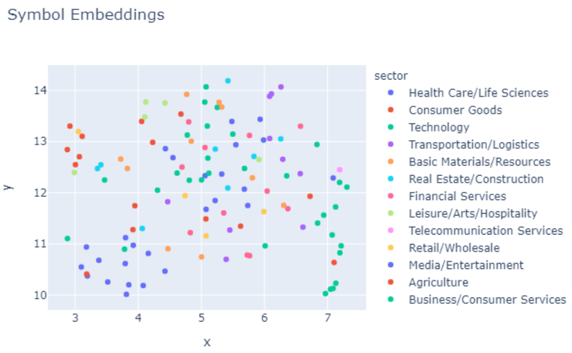

Fig. 1: Company embeddings (Reduced with UMAP [16]) are plotted and colored by sector. Same-colored groups show similarities found between companies in the same sector.

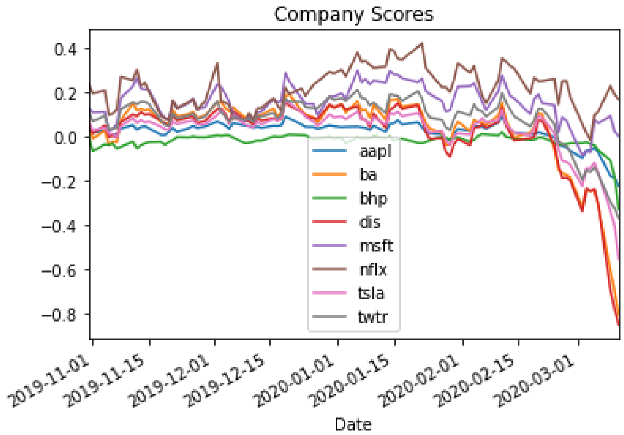

Fig. 2: The relevant sentiment (RS) scores for several companies during the onset of COVID-19 in the United States. While Disney's (DIS) RS drops, Netflix's (NFLX) RS rises.

A naive trading strategy which trades solely based on relevant sentiment (buy if positive, sell if negative) was also backtested a few months and showed significant gains over other simple strategies such as Cross Moving Average or long holding SPY. The duration of the backtest was not long enough to conclusively know if the strategy is good and this will be further tested.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[2] M. Velay and F. Daniel, "Using NLP on news headlines to predict index trends," *CoRR*, vol. abs/1806.09533, 2018.

[3] K. Joshi, B. N, and J. Rao, "Stock trend prediction using news sentiment analysis," *International Journal of Computer Science and Information Technology*, vol. 8, pp. 67–76, 06 2016.

[4] D. Shah, H. Isah, and F. H. Zulkernine, "Predicting the effects of news sentiments on the stock market," *CoRR*, vol. abs/1812.04199, 2018.

[5] H. Mao, S. Counts, and J. Bollen, "Predicting financial markets: Comparing survey, news, twitter and search engine data," 2011.

[6] L. D. Corro and J. Hoffart, "Unsupervised extraction of market moving events with neural attention," 2020.

[7] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using structured events to predict stock price movement: An empirical investigation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1415–1425, Association for Computational Linguistics, Oct. 2014.

[8] H. Liu, "Leveraging financial news for stock trend prediction with attention-based recurrent neural network," 2018.

[9] M. Levene, "Market trend prediction using sentiment analysis: lessons learned and paths forward," 06 2018.

[10] F. Ming, F. Wong, Z. Liu, and M. Chiang, "Stock market prediction from wsj: Text mining via sparse matrix factorization," in *2014 IEEE International Conference on Data Mining*, pp. 430–439, Dec 2014.

[11] Z. Hu, W. Liu, J. Bian, X. Liu, and T.-Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," 12 2017.

[12] *hiQ Labs, Inc. v. LinkedIn Corp.* 2019.

[13] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014.

[14] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," 01 2015.

[15] T. De Smedt and W. Daelemans, "Pattern for python," *J. Mach. Learn. Res.*, vol. 13, p. 2063–2067, June 2012.

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.