

Market Efficiency of News Events

Adrian Beer

Contents

1	Introduction	1
1.1	Categories	1
1.1.1	Labeling	1
2	Related Work	2
3	Methodology	3
4	Labeling	3

1 Introduction

One possible explanation is that "low-capitalisation, young, unprofitable, low-dividend-paying, high-volatility and highgrowth companies are difficult to arbitrage or value according to traditional financial theory and are therefore very sensitive to investor sentiment".

1.1 Categories

Peng and Jiang (2015): Bag of Keywords, Polarity Score, Category Tagging... Using initial seed words for each category. Categories: new-product, acquisition, pricerise, price-drop, law-suit, fiscal-report, investment, bankrupt, government, analyst-highlights.

1.1.1 Labeling

- Next Days Close
- Next Close/Open
- Trailing Stop-Loss

2 Related Work

Salbrechter (2021) investigates the impact of financial news on the daily returns of S&P 500 constituents using a self-fine-tuned BERT model. Their algorithm produce the CLS Token from FinNewsBert and a Topic vector from Text2Topic and feed these into a standard feed forward NN, which then produces probabilities for the 3 classes. They find that the news is priced in within one day, but not instantaneously. Their FinNewsBert model is retrained every two years in order to obtain a large time-frame for their out-of-sample study. They consider the problem of look-ahead-bias of the base BERT model, due to the data that it is trained on. They use a dataset of financial news published by Refinitiv from 1996 to 2020 and from Refinitive Datastream daily price data of 1330 companies, which were at some point in time S&P constituents. As part of the preprocessing they convert news articles to lower case, remove all numbers, punctuation marks and brackets, so that only letters remain. Multiple news articles about one company published in quick succession are merged into a single document. (Q: How is this handled coupled with the token limitations of BERT?) An algorithm is used to discern fresh and stale news. The labels for the training of FinNewsBert are constructed based on the idiosyncratic price movement following the news, namely on the following formula:

$$IR_{i,t} = R_{i,t} - R_{f,t} - \beta_{i,t} * (R_{S\&P500,t} - R_{f,t})$$

Afterwards z-scores are calculated for each stock, in order to avoid over-weighting price movements in smaller, more volatile stocks.

$$z_{i,t} = \frac{IR_{i,t} - \mu_{i,t}}{\sigma_{i,t}}$$

They find improved classification performance, with the topic vector being used as input. With the Text2Vec algorithm they differentiate between analyst forecasts, earnings reports, monetary policy and business/strategic.

Ke et al. (2020) propose a text-mining methodology, named SESTM, to extract sentiment information from text. They used a bag-of-words representation of news articles. To distinguish fresh from stale news a measure of novelty was constructed based on the cosine similarity of an articles with all other articles about the same firm five trading days prior.

$$\text{Novelty}_{i,t} = 1 - \max_{j \in \mathcal{X}_{i,t}} \left(\frac{d_{i,t} \cdot d_j}{\|d_{i,t}\| \|d_j\|} \right)$$

They find that price responses to news are larger for fresh news. They also found larger price responses for smaller and also for more volatile stocks. Price movements for more volatile stocks last up to three days on average, while news about low volatility stocks are incorporated after just one day of trading.

Liu and Stentoft (2023) analyze the predictability of stock returns for different time horizons, ranging from 1-min to 30-min, and market sectors between 2005 and 2016. They find high profitability even after transaction costs. Intraday predictability decreased with higher time horizons

These results are consistent with the observations of Baker (2007b), who observes that low-capitalisation, young, unprofitable, low-dividend-paying, high-volatility and highgrowth companies are difficult to arbitrage or value according to traditional financial theory and are therefore very sensitive to investor sentiment.

3 Methodology

Because pre-news price movements have shown to correlate with the sentiment of the news, we incorporate this information into our feature set Ke et al. (2020).

4 Labeling

Adjusting returns to the market is important. Salbrechter (2021):

$$IR_{i,t} = R_{i,t} - R_{f,t} - \beta_{i,t} * (R_{S\&P500,t} - R_{f,t})$$

Chen (2021):

$$r_{s,t} = \frac{P_{s,t+\Delta t}}{P_{s,t}} - \frac{P_{m,t+\Delta t}}{P_{m,t}}$$

Since news can take from 1 day up to 3 days to be fully incorporated in the stock price, ideally we set our label accordingly. However the longer this time horizon, the higher the probability of having a second news event in this period. How big is this problem??

References

- Qinkai Chen. Stock Movement Prediction with Financial News using Contextualized Embedding from BERT, July 2021. URL <http://arxiv.org/abs/2107.08721>. arXiv:2107.08721 [cs, q-fin].
- Zheng Tracy Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting Returns with Text Data, September 2020. URL <https://papers.ssrn.com/abstract=3389884>.
- Fred Liu and Lars Stentoft. Intraday Stock Predictability Everywhere. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4496917. URL <https://www.ssrn.com/abstract=4496917>.
- Yangtuo Peng and Hui Jiang. Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks, June 2015. URL <http://arxiv.org/abs/1506.07220>. arXiv:1506.07220 [cs] version: 1.
- Stefan Salbrechter. Financial News Sentiment Learned by BERT: A Strict Out-of-Sample Study, November 2021. URL <https://papers.ssrn.com/abstract=3971880>.