Faculty of Computer Science

Master of Data Science

Moscow
2022

Forecasting a trend reversal in the movement of share price using news flow analysis

Supervisor: Candidate of Sciences (PhD) in Mathematical Modelling, Programme Academic Business Analytics and Big Data Systems, Armen Beklaryan

Student: Andrei Li

The objective, process and result of this research

Objective

- The goal of this paper is to investigate the applicability of Natural language processing methods to analyze news flow and predict the trend reversal of stock prices for a particular company.

Process

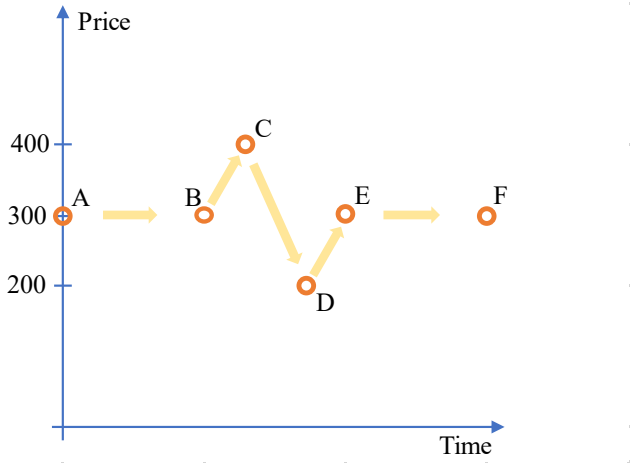
- A review of scientific papers in this field was conducted.
- Two promising approaches within sentiment analysis were chosen as the main ones.
- These approaches were tested on a text stream of news about public companies.

Result

- It was shown the co-dependence between the change of companies' value and the change of sentiment estimates of the news background within a day with R^2 at the level up to 0.09-0.13 by two independent approaches – this may be probably a new scientific result.

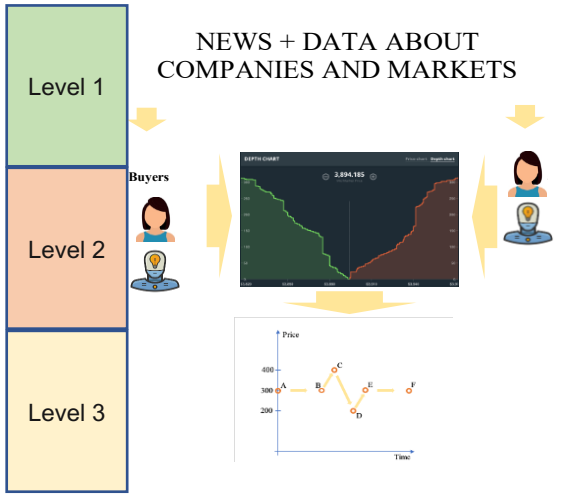
Formulation of the problem

1 Why exactly is a trend reversal important?



A precise prediction of trend reversal allows to earn the profit on price changes, for this example: BC + CD + DE = 100 + 200 + 100 = 400.

2 Why exactly news flow analysis is important for a trend reversal?



The news flow is probably the top-level factor determining the landscape for the factors related to order book depth or time series analysis. With some assumptions, it may lead to the steepest trend reversals. And it may be an easier prediction task, since they are pronounced, obvious.

Review of the available results and current approaches

1 Method

For scientific articles
scholar.google.com & library.hse.ru.

For books
books.google.com & others.

Keywords: "news", "stock", "price", “sentiment”.

2 Interesting general observations

Relevant sources have been founded by scientific areas and years			
	1982-1999	**2000-2009**	**2010-2022**
Mathematics	13%	0%	9%
Economics	87%	100%	18%
Computer Science	0%	0%	73%
Total number	**14**	**12**	**11**

Most of Computer Science works use Accuracy instead of R-squared or RMSE. Some of them state Accuracy >0.6. Both this points may look a bit suspicious.

3 Main review results

- The main factors for news flow analysis, for future consideration, were selected the nature of the news and the degree of dissemination of the news.
- The most common described approach (2010-2022) for the news nature evaluation are dictionary-based sentiment analysis.

Programme techniques and methods used: https://github.com/liandreigithub/thesis

	Observer or/and tested	Method of selection	**Finally applied**
Stock data source	API: Alpha vantage; Twelve data; Exante; Tinkoff; Moex iss; Yahoo Finance	Number of criteria evaluation	**Tinkoff API**
News data source	API: Bloomberg, Google news, News API, EOD Historical Data, Finnhub API	Number of criteria evaluation	**Finnhub API**
	Web-scraping: Selenium library for Bloomberg, Investing.com, Google news	Brute force	**-**
General computations and statistic analysis	Libraries: pandas, numpy, matplotlib, seaborn, datetime, sklearn, xgboost, nltk, torch, transformers, pysentiment2	-	**Pandas, NumPy, Matplotlib, Seaborn, Datetime, Sklearn, XGBoost, NLTK, Torch, Transformers, Pysentiment2**
Sentiment analyses	Dictionary-based: VADER; TextBlob; Loughran-McDonald financial sentiment word list; Harvard sentiment dictionary	Compliance of results with theoretical prerequisites	**VADER; Loughran-McDonald financial sentiment word list; Harvard sentiment dictionary**
	Bidirectional Encoder Representations from Transformers: BERT, FinBERT	Brute force	**FinBERT**

Experiments: data samples

	Dictionary-based approach			Dictionary-based approach, BERT-based approach
	TOP-10 companies cap. sample	TOP-100 (91) companies cap. sample	TOP-4 companies cap. sample	TOP-1 companies cap. sample
Time period	Q1 2021- Q1 2022 (Finnhub API limitation for news)			
Time scale	day	day	day	hour
Observation sample criteria	Days with 5 to 10 and 90 to 95 percentiles of trend reversal indicator.	The day corresponding to the 5th percentile and the day corresponding to the 95th percentile of trend reversal indicator.	Days corresponding to 5-15 percentile, 45-55 percentile, and days corresponding to 85-95 percentile of trend reversal indicator.	
Number of observed days per company	20-25	2	65-85	209 => only 3 of 4 days per week are awaliable to calculate trend reversal indicator => 160 (74 training days, 84 test days).
Number of news per company per day	45-55	10-15	30-40	50-60
Objective	**Exploring data, computation time-scale.**	**Showing significance of sentiment indicator for value up and down days.**	**Rank match check: days with value up, neutral and down match high, medium and low scores of sentiment indicator.**	**Regression analysis**
Nature of the news/sentiment indicator	If there are more positive words than negative - we consider the news itself to be positive. Then for each day we calculate the share of positive news in the total news stream and finally determine the share of positive news in the total news stream for the average rise and fall day.			Average negative words proportion in news text.
Trend reversal indicator	On today's day, determine the average price value (from the daily median values) for the preceding five days of trading.			
News dissemination indicator	News count per day.			

Experiments: Implementation and results of the dictionary-based approach

1 First three iterations: TOP-10, TOP-100 and TOP-4 companies by market capitalization

Objective: determine the scale for the timing of the calculations; try find out a difference in sentiment marks between markedly up or markedly down days; evaluate applicability of 4 sentiment dictionary libraries.
Process on data: graphical analysis; Shapiro-Wilk test; Two sample t-test with unequal variances; rank correspondence.

Key results: Sentiment indicators for VADER; Loughran-McDonald; Harvard libraries are in line with the theoretical assumptions for all three samples. On days when company stocks rise noticeably (a trend reversal), there is a low level of negative news, and the opposite is true. The difference is significant.

2 Forth iteration: TOP-1 company by market capitalization

Objective: calculation of the possible degree of influence of the nature of the news flow and the degree of news spread on the relative price changes (trend reversal) within a day.
Process on data: regression analysis for general linear and logarithmized form of Cobb-Douglas function.

Key results: degree of news spread may not be the significant factor for trend reversal forecasting. There is a possible co-dependence between the change of companies' value and the change of sentiment estimates of the news background within a day with R-squared at the level up to 0.133, adjusted R-squared is 0.097.

Experiments: Implementation and results of the BERT (FinBERT) approach

1 Choosing a classification model (BERT vs. FinBERT)

Objective: get sentiment classification model.
Process on data: attempt to train BERT model; classification of test days with BERT and FinBERT.
Key results:

BERT: error matrix

True sentiment \ Predicted sentiment	neutral	positive	negative
neutral	56	0	0
positive	29	0	0
negative	29	0	0

FinBERT error matrix

True sentiment \ Predicted sentiment	neutral	positive	negative
neutral	42	8	6
positive	12	14	3
negative	4	10	15

2 Regression analysis with pretrained classification model (FinBERT)

Objective: calculation of the possible degree of influence of the nature of the news flow and the degree of news spread on the relative price changes (trend reversal) within a day.
Process on data: regression analysis (Linear, Random Forest, XGBoost).
Key results:

Based on the news headers:

	Linear	Random Forest	XGBoost
R^2	0.032	0.101	0.119
RMSE	2.011	1.876	1.857

Based on the news descriptions:

	Linear	Random Forest	XGBoost
R^2	0.012	0.098	0.078
RMSE	1.966	1.879	1.900

Experiments: Predicting the nature of the news flow at the end of the day

Objective: an attempt to predict the stock price change on the test data.

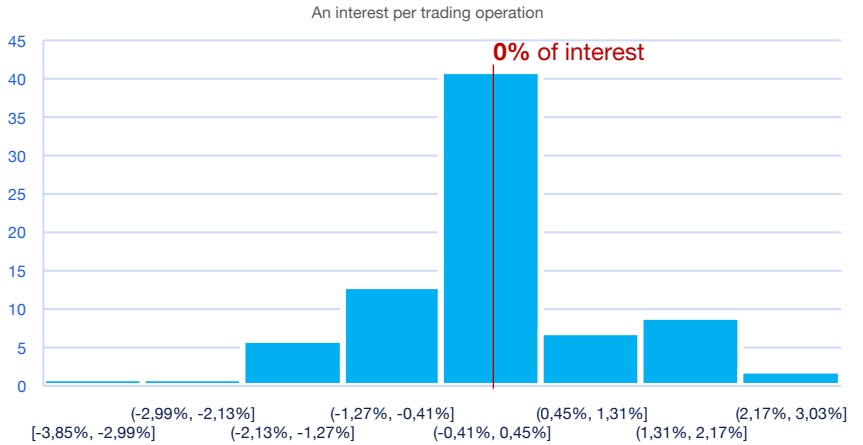
Process on data:

1. For each hour of the day, we calculate the number of published news items and the average proportion of negative words in these news items.

2. If there are more than 7 news items published, and if the nature of the news flow is very good or very bad (more than the 75th percentile or less than the 25th percentile for the proportion of negative words in the news per day in the training sample), we would decide to buy or sell the stock at that hour.

3. Then we would close the position at the end of the day, for example within the last three trading hours. The difference in price between the time we made the decision and the time we closed the position at the end of the day is our interest.

Key results:



The total interest of operations for all days on the test sample was 1.11%, the median for interest per operations was 0.00%, while the standard deviation for interest per operations was 1.06% for 84 test observations.

Conclusion

According to the results of approbation: the correlation between the level of companies' capitalization and the number of daily published news for each of them was confirmed; it was shown the **co-dependence between the change of companies' value and the change of sentiment estimates of the news background within a day with R^2 at the level up to 0.09-0.13 by two independent approaches – this may be probably a new scientific result**, because the presentation and measurement of R^2 distinguishes this work from the main scientific studies reviewed, which use and show the accuracy metric.; the proposed method for predicting the trend reversal of share price within a day showed no significant predictive ability.

The main promising direction for the development of this work is a meticulous analysis in terms of:

1. Minimization of time scale detail for tracking the timing of news publication and stock price changes;
2. Using a mixed approach based on dictionary-based libraries and Bidirectional Encoder Representations from Transformers.
3. Ranking the news sources by significance.