NLP For Sentiment Analysis in Financial Market

Ruixuan Chen

Introduction

Sentiment analysis for financial market has been given great hopes as the Natural Language Processing (NLP) technique is mature enough to support the application with high confidence. However, due to the complexity of the financial market, it's still a challenging task to generate sound and rational signals. The problems come from the data and the algorithm. It is particularly important for sentiment data as the algorithm should be designed in accordance with the intrinsic logic of this specific data type that differs from traditional factors. In this project, I tried to build a framework to tackle the mentioned problems with some toy data and did the backtest on quantconnect[1].

Data & Pre-processing

The data were from Kaggle and the Internet

The purpose is to extract all the companies in the news while maintaining the mapping relation for the input of target-dependent BERT (TD-BERT[2]). The key is that a single news may contain multiple companies and a company may occur several times.

- Names2Ticker (mapping from company's aliases to ticker symbol)

	Exchange	Symbol	Shortname	Longname	Sector
0	NMS	AAPL	Apple Inc.	Apple Inc.	Technology
1	NMS	MSFT	Microsoft Corporation	Microsoft Corporation	Technology
2	NMS	GOOG	Alphabet Inc.	Alphabet Inc.	Communication Services
3	NMS	GOOGL	Alphabet Inc.	Alphabet Inc.	Communication Services
4	NMS	TSLA	Tesla, Inc.	Tesla, Inc.	Consumer Cyclical

Fig. 1 Ticker symbol with company's aliases and Sector sample[3]

- CEO2Ticker (mapping from company's CEO to company's name then to ticker symbol)

Rank	Chief Executive Officer	Company	Country
1	Tim Cook	Apple	US
2	Satya Nadella	Microsoft	US
3	Sundar Pichai	Alphabet (Google)	US
4	Andy Jassy	Amazon	US
5	Elon Musk	Tesla	US
6	Mark Zuckerberg	Meta (Facebook)	US
7	Warren Buffett	Berkshire Hathaway	US

Fig. 2 CEO with company's name sample[4]

With the two mappings above, applying the following logic with SpaCy

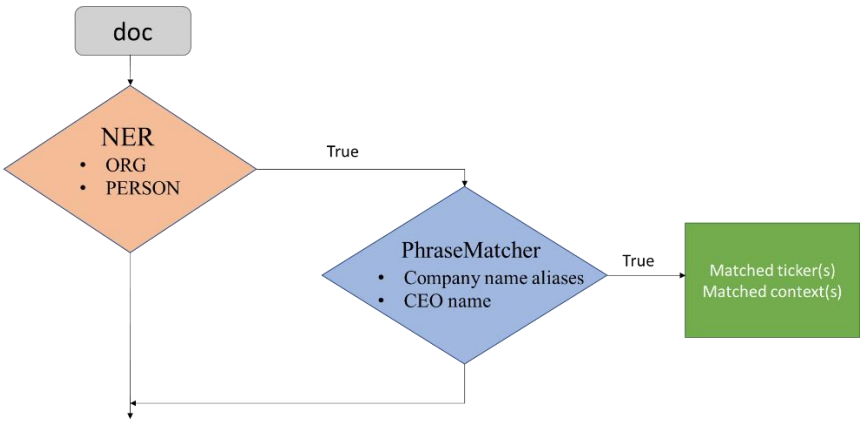

Fig. 3 Logic flow to screen the news with relevant context

The news data with target-dependent company's title level information is summarised below. Matches are the keyword that fit into our screening process and the tickers are the

final sentiment analysis objects we wish to attribute to. For this dataset (~140,000 in total), ~2.9% (4,047) news titles were extracted with 119 different tickers.

	id	title	publication	content	date	tickers	matches
3031	176619	U.S. blocks health insurer Aetna's $34 billion...	Reuters	The ruling is another victory for the U. S. Ju...	2017-01-23	['HUM']	['Humana']
1147	65665	Microsoft CEO Satya Nadella says Bill Gates' o...	Business Insider	'' 'Around 1980, Bill Gates gave Microsoft, th...	2017-02-22	['MSFT', 'MSFT']	['Microsoft', 'Satya Nadella']
2385	130565	Banker places odds on Apple's next acquisition...	New York Post	Wall Street's about Apple's merger and acqui...	2017-05-05	['AAPL']	['Apple']
3295	185056	Dow, S&P off to worst four-day Jan start ever ...	Reuters	China allowed the biggest fall in its yuan cur...	2016-01-07	['DOW']	['Dow']
1465	69393	Berkshire Hathaway's legendary annual meeting ...	Business Insider	"Berkshire Hathaway's annual meeting is over."...	2016-05-01	['BRK-B']	['Berkshire Hathaway']

Fig. 4 Title level news screening sample[5]

Applying TD-BERT inference we have the sentiment score and corresponding probabilities.

	id	title	publication	date	tickers	matches	neutral	positive	negative	score
984	23195	Twitter Appoints Debra Lee, Adding Diversity t...	New York Times	2016-05-17	['TWTR']	['Twitter']	[0.7272063493728638]	[0.25229039788246155]	[0.02050325833261013]	[-1]
2223	193002	Nasdaq names Friedman CEO; Greifeld to be chai...	Reuters	2016-11-14	['NDAQ']	['Nasdaq']	[0.9650331735610962]	[0.024829719215631485]	[0.010137098841369152]	[-1]
725	69063	Mark Zuckerberg has a 'yellow' version of Face...	Business Insider	2016-04-10	['META', 'META']	['Mark Zuckerberg', 'Facebook']	[0.4139641225337982, 0.7223705053329468]	[0.45506638288497925, 0.14487498998641968]	[0.13096946477890015, 0.13275445997714996]	[-1, -1]
3450	129182	Suspect arrested in Google exec's murder	New York Post	2017-04-15	['GOOG']	['Google']	[0.49776455760002136]	[0.007179925683885813]	[0.49505552649497986]	[1]
2345	72809	Walmart's new gas-and-grocery hybrids could be...	Business Insider	2016-12-02	['WMT', 'AMZN']	['Walmart', 'Amazon']	[0.32328009605407715, 0.7665201425552368]	[0.14681944251060486, 0.03740778937935829]	[0.5299004912376404, 0.19607211649417877]	[1, 1]

Fig. 5 TD-BERT sentiment score[6]

Visualisation

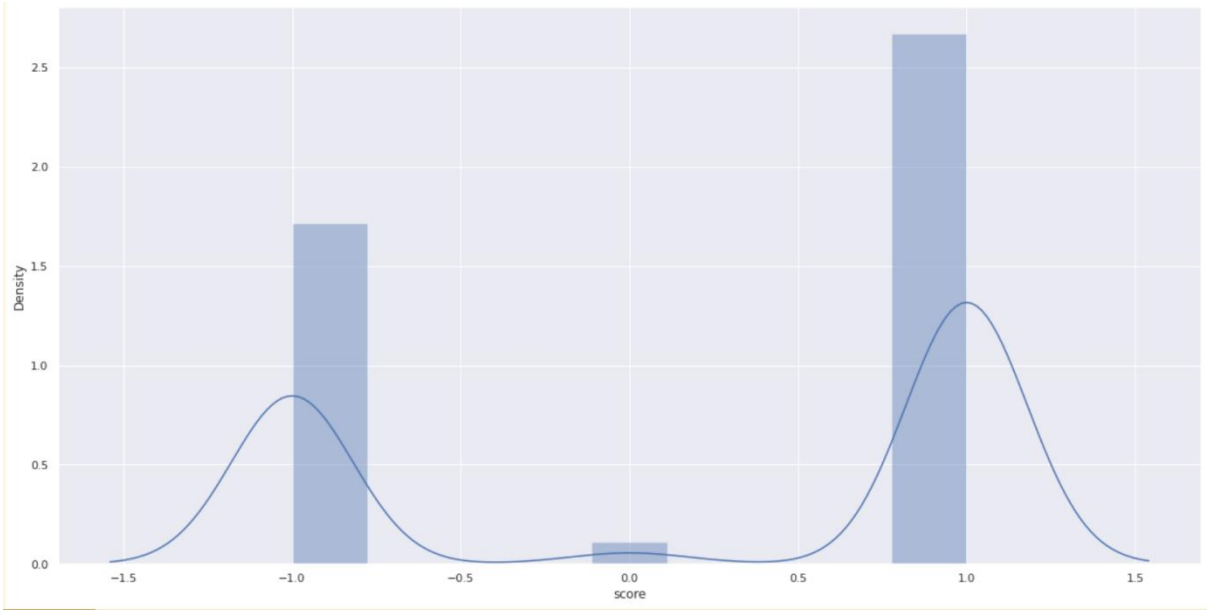

Fig. 6 Overall Sentiment Score distribution

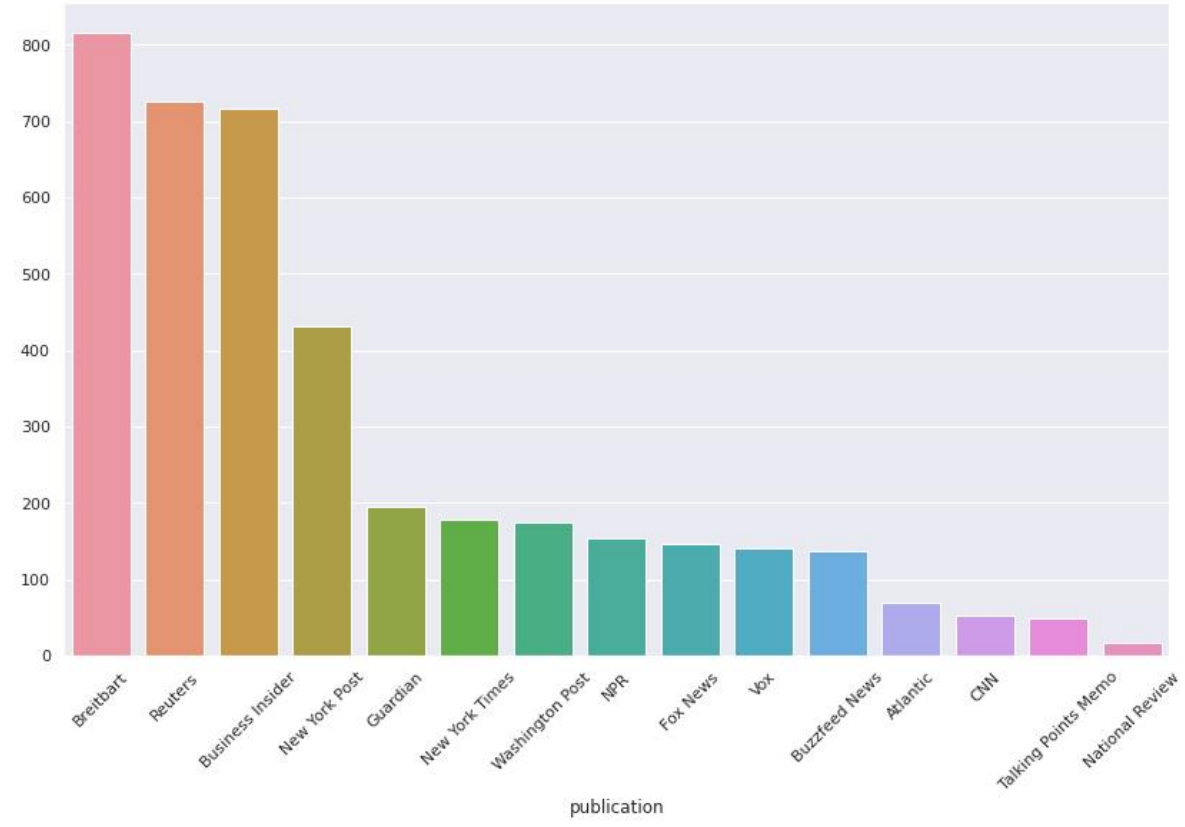

Fig.7 News Count Distribution vs. media

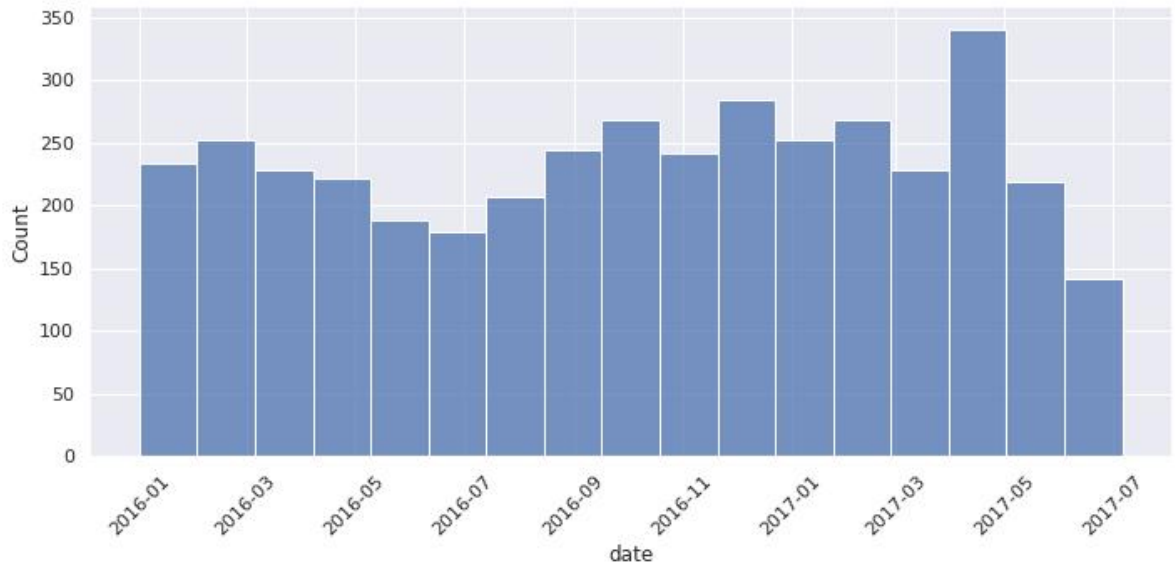

Fig. 8 News Count Distribution vs. time

Discussion

Consider the Gibbs inequality:

$$-\sum_i p_i log(p_i) \leq -\sum_i p_i log(q_i)$$

To increase the amount of information from we can adjust our information set from p_i to q_i. However, it only guarantees a greater load of information if we stand on the set of q_i but not necessarily the signal to noise ratio. Then an intuitive idea is that the modified set q_i should carry more perceived truth or consensus about the market in our context. Under such circumstance, we could have not only more information but also higher resolution. And given the wisdom of the crowd, the new set q_i is constructed via the overall sentiment from all the media.

According to Bayesian formula,

$$P(pos|media_k) \propto P(media_k|pos)P(pos)$$

P(pos) or P(neg) can be estimated from the overall media positive/negative rate from the past. The intuition is that we expect the ensemble sentiment should be “fairer”, and the ensemble should have lower variance. Given the observed posterior P(pos|media), we also need to estimate the likelihood P($media_k$|pos) which is the w_k. w_k could be a function (Bayesian statistics but for simplicity just make it a number)

$$w_k = \frac{P(pos|media_k)}{P(pos)}$$

For negative sentiment the w_k is negative, and the aggregated result for a single stock will be the linear combination of the positive and negative weight.

Applying the w_k, we have the following:

	date	publication	tickers	score	sector	pos_weight	neg_weight	media_weight
0	2016-01-01	Breitbart	TWTR	1	Communication Services	0.834521	-1.446227	0.834521
1	2016-01-03	Breitbart	TWTR	-1	Communication Services	0.834521	-1.446227	-1.446227
2	2016-01-04	Breitbart	META	-1	NaN	0.834521	-1.446227	-1.446227
3	2016-01-04	Breitbart	AMZN	-1	Consumer Cyclical	0.834521	-1.446227	-1.446227
4	2016-01-05	Breitbart	DOW	1	Basic Materials	0.834521	-1.446227	0.834521

Fig. 9 Corrected weight sample

The aggregated weight tells how informative our data is, and our strategy for investment is based on the information intensity. We need to treat different news with different methods indicated by the aggregated weight we have acquired. For those with a very high absolute aggregated weight, it means the news is fresh and breaking, and we follow that trend. If it is the other way around, it can be interpreted that the market has digested the news, and we expect a reversion. Between these two cases, it implies that the market is partially pre-acted, and we are in the middle of the exponentially decaying signal, therefore we will still follow the residual trend.

The framework above is for a single stock. It is similar for a portfolio but with much more technical details such as the weight allocation constrained not only by Sharpe ratio but also the aggregated weight.

Here was the backtest result with the strategy I came up with conditioned on different parameters. The test period was 01/01/2016 – 06/07/2017 and the stock was Amazon (AMZN).

Fig. 10 AMZN buy and hold

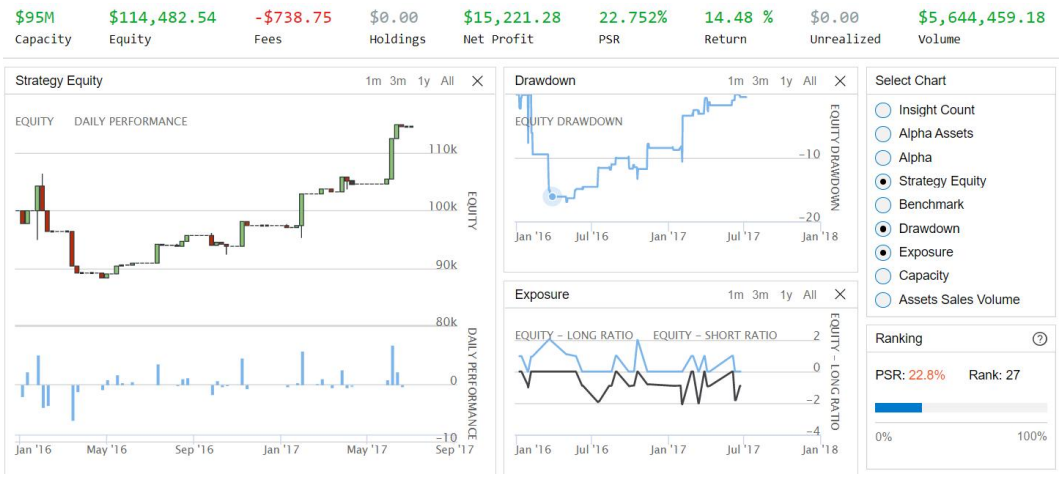

Fig. 11 AMZN with random parameters

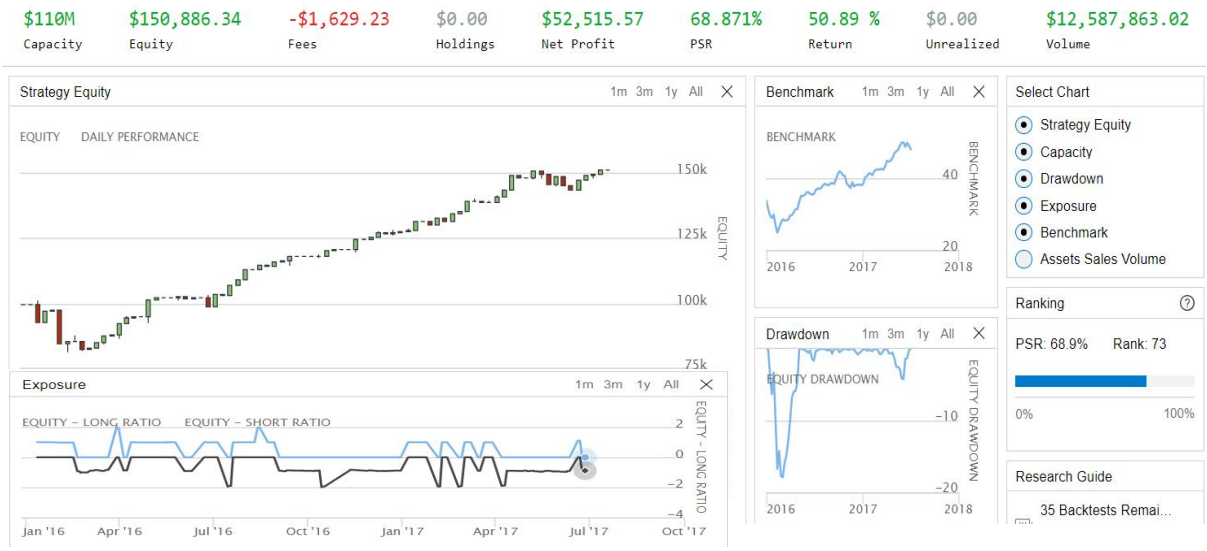

Fig. 12 AMZN with optimised parameters

It is very likely that the strategy will excel considering the return and low correlation with the market. As is observed in Fig. 12, from April 16th it shows consistent green bars with little red ones while taking both long and short positions. This is also true for Fig. 11 even though it has a much smaller return due to unoptimised parameters.

Conclusion

In this project, I applied NLP techniques to extract sentiment from news title about the stock market and designed a bespoke framework for investment strategy based on the data. The backtest result showed that the framework would function well if proper parameters could be given. This will rely on the deeper understandings of the data and the parameters. And there are tremendous amount of details could be improved to have a more reliable and better result. It includes but not limits to the NER in sentiment extraction, rational separation of strategy selection and parameter optimisation.

Reference

1. https://www.quantconnect.com/
2. NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles (Hamborg & Donnay, EACL 2021)
3. https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?select=sp500_companies.csv
4. https://ceoworld.biz/2022/01/25/the-worlds-most-influential-ceos-and-business-executives-of-2022/
5. https://www.kaggle.com/datasets/snapcrack/all-the-news
6. https://huggingface.co/fhamborg/roberta-targeted-sentiment-classification-newsarticles/blob/7a789d8a34f44a3607072ceee514d15f7fead3bf/README.md