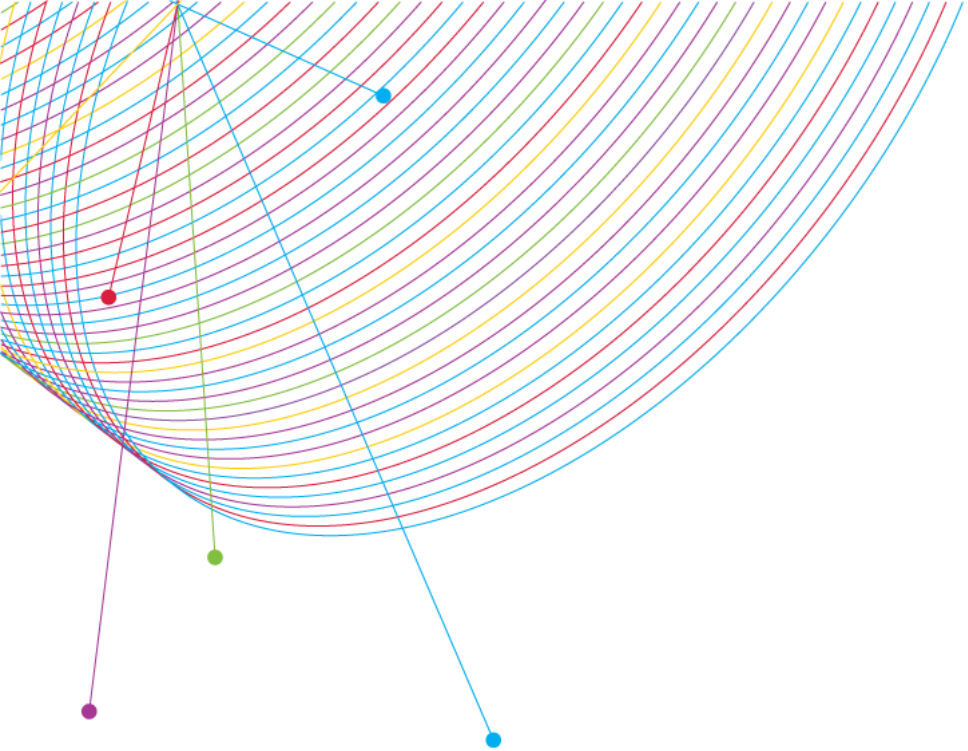


Predicting Stock Return with Text Data

RECENT NEWS

Recent breaking news all over the world

Africa

- ‘This is huge’: Locust swarms in Africa are worst in decades

Australia

- Australia fires: The fires are still burning. And they’ll be burning for months to come

China

- The coronavirus from China is new, and that makes everything dicier

US

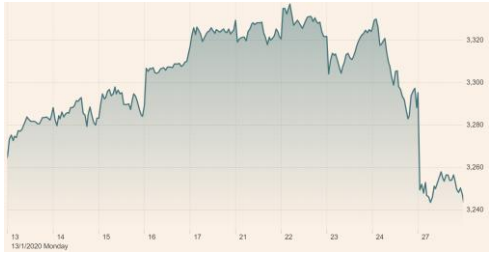
- Kobe Bryant, daughter killed in copter crash, 7 others dead

HOW IS THE MARKET

Review of the current market

Financial Market

- **S&P 500 Index (01/13/2020—01/27/2020):**

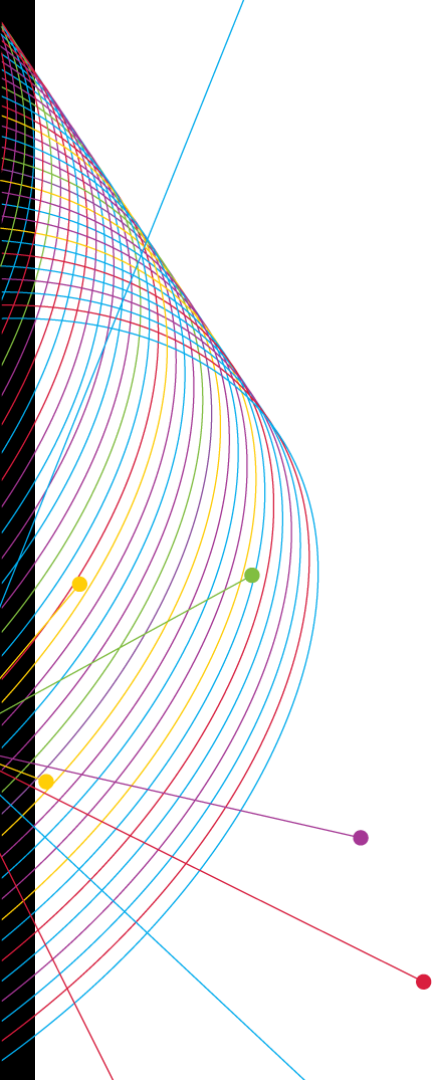


- **VIX Index (01/02/2020 – 01/27/2020):**

Theories

- **Keynes (1936): Animal Spirits**
- **Eugene Fama (1970): Efficient Market Hypothesis**
- **Shiller (2002): Feedback Loop & Attention Cascade**

Contents:



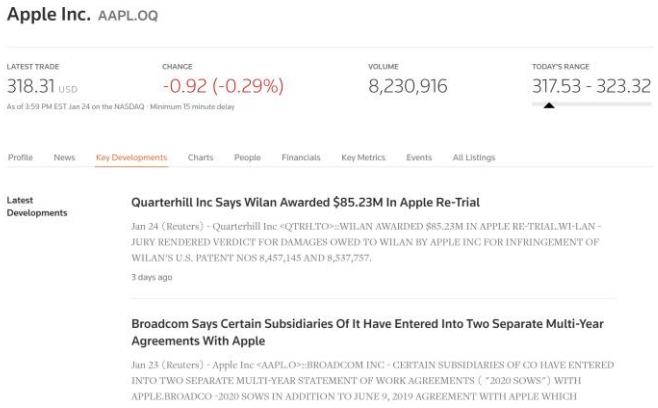
PART 1. DATA

DATA MINING

Scratching all the news of the stocks in NYSE, NASDAQ and ASE

Tickers

- Download all the tickers from NYSE, NASDAQ, and ASE;
- Search for each ticker's website on Thomson Reuters;



Selenium & Scrapy

- Use Python **Selenium** and **Scrapy** to scratch 200,000+ news articles;
- Utilize **NLTK** and **Beautifulsoup** package in Python to conduct text pre-processing;

Selenium

Scrapy

Natural Language Analyses with NLTK

Beautifulsoup

TEXT PRE-PROCESSING

Using NLTK package to clean the news we got by the web crawler

- The raw news (example):

```
 German biotech group Morphosys seeks European partner for debu
t drugNovartis said on Thursday it will pay 95 million euros
($110.49 million) to Galapagos and MorphoSys to license a prosp
ective medicine targeting the skin condition atopic dermatitis,
as the Swiss drugmaker strengthens its foothold in immunology.
```

- Converting text to lowercase:

```
 german biotech group morphosys seeks european partner for debu
t drugnovartis said on thursday it will pay 95 million euros
($110.49 million) to galapagos and morphosys to license a prosp
ective medicine targeting the skin condition atopic dermatitis,
as the swiss drugmaker strengthens its foothold in immunology.
```

- Regular expression:

```
 german biotech group morphosys seeks european partner for debu
t drugnovartis said on thursday it will pay     million euros
million  to galapagos and morphosys to license a prospective me
dicine targeting the skin condition atopic dermatitis  as the s
wiss drugmaker strengthens its foothold in immunology
```

- Tokenizing:

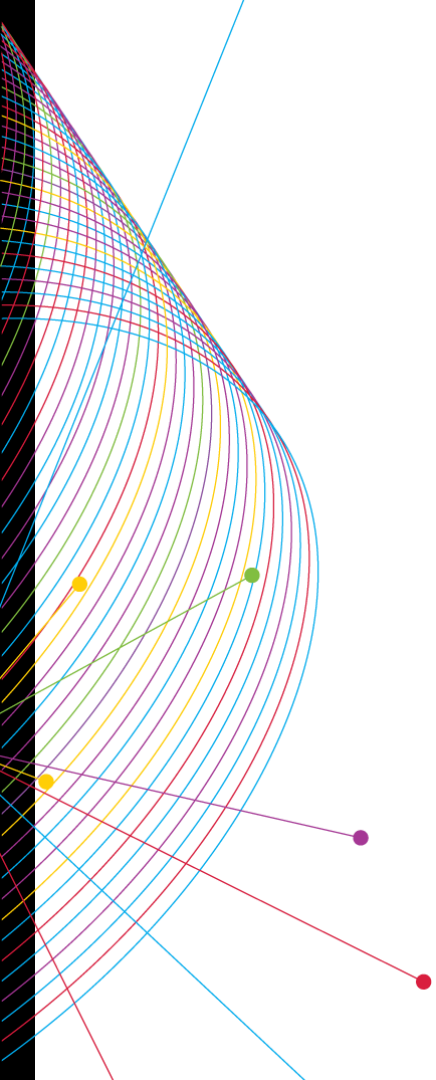
```
['german', 'biotech', 'group', 'morphosys', 'seeks', 'europea
n', 'partner', 'for', 'debut', 'drugnovartis', 'said', 'on', 't
hursday', 'it', 'will', 'pay', 'million', 'euros', 'million',
'to', 'galapagos', 'and', 'morphosys', 'to', 'license', 'a', 'p
rospective', 'medicine', 'targeting', 'the', 'skin', 'conditio
n', 'atopic', 'dermatitis', 'as', 'the', 'swiss', 'drugmaker',
'strengthens', 'its', 'foothold', 'in', 'immunology']
```

- Deleting Stop words:

```
['german', 'biotech', 'group', 'morphosys', 'seeks', 'europea
n', 'partner', 'debut', 'drugnovartis', 'said', 'thursday', 'pa
y', 'million', 'euros', 'million', 'galapagos', 'morphosys', 'l
icense', 'prospective', 'medicine', 'targeting', 'skin', 'condi
tion', 'atopic', 'dermatitis', 'swiss', 'drugmaker', 'strengthe
ns', 'foothold', 'immunology']
```

- Stemming & Lemmatizing:

```
['german', 'biotech', 'group', 'morphosi', 'seek', 'european',
'partner', 'debut', 'drugnovarti', 'said', 'thursday', 'pay',
'million', 'euro', 'million', 'galapago', 'morphosi', 'licens',
'prospect', 'medicin', 'target', 'skin', 'condit', 'atop', 'der
mat', 'swiss', 'drugmak', 'strengthen', 'foothold', 'immunolo
g']
```



PART 2. METHODOLOGY

MODEL SETUP

Create the word matrix with the sentiment-charged words for the different news

Word Matrix

- Record the word (or phrase) counts of the certain article in a vector;
- Combine those vectors to form the word matrix.

$$D = [d_1 \quad \dots \quad d_n]'$$

$$\begin{bmatrix} d_{1,1} & \cdots & d_{1,m} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & d_{n,m} \end{bmatrix}$$

Sentiment-charged words

- Calculate the frequency with which word j co-occurs with a positive return;

$$f_j = \frac{\#\text{ articles including word } j \text{ AND having } \operatorname{sgn}(y) = 1}{\#\text{ articles including word } j}$$

- Set the threshold values to filter the sentiment-charged words;

$$\widehat{S} = \{j : f_j \geq 1/2 + \alpha_+, \text{ or } f_j \leq 1/2 - \alpha_-\} \cap \{j : k_j \geq \kappa\}.$$

SENTIMENT DISTRIBUTION LEARNING

Based on the Multinomial Distribution, estimate the crucial estimators, O+ and O-

Preparation

- Distribution assumption: sentiment-charged word counts are generated by a mixture multinomial:

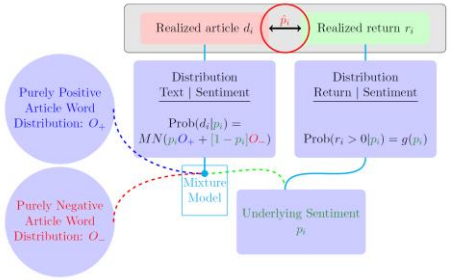
$$d_{i,[S]} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1-p_i)O_-\Big)$$

- P value setup: for each news, we assign a sentiment value p based on the rank of the relevant stock's return:

$$\widehat{p}_i = \frac{\text{rank of } y_i \text{ in } \{y_l\}_{l=1}^{n}}{n}$$

Distribution Estimation

- By setting up the sentiment values for each news, to estimate the estimators, O+ and O- (two topic vectors):



F: a vector of frequency

$$F = \frac{1}{2}(O_+ + O_-)$$

T: a vector of tone

$$T = \frac{1}{2}(O_+ - O_-)$$

ADDITIONAL CALCULATION PROCESS

With the result of O+ and O-, add a penalty term to optimize the calculation process

O+ & O-

- Calculation process:

$$\mathbb{E}\widetilde{d}_{i,[S]} = \mathbb{E}\frac{d_{i,[S]}}{s_i} = p_i O_+ + (1-p_i)O_-$$

$$\mathbb{E}\widetilde{D}' = OW, \quad \text{where} \quad W = \begin{bmatrix} p_1 & \cdots & p_n \\ 1-p_1 & \cdots & 1-p_n \end{bmatrix}, \quad \text{and} \quad \widetilde{D} = [\widetilde{d}_1, \widetilde{d}_2, \ldots, \widetilde{d}_n]'$$

$$\widetilde{O} = \widehat{D}\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$$

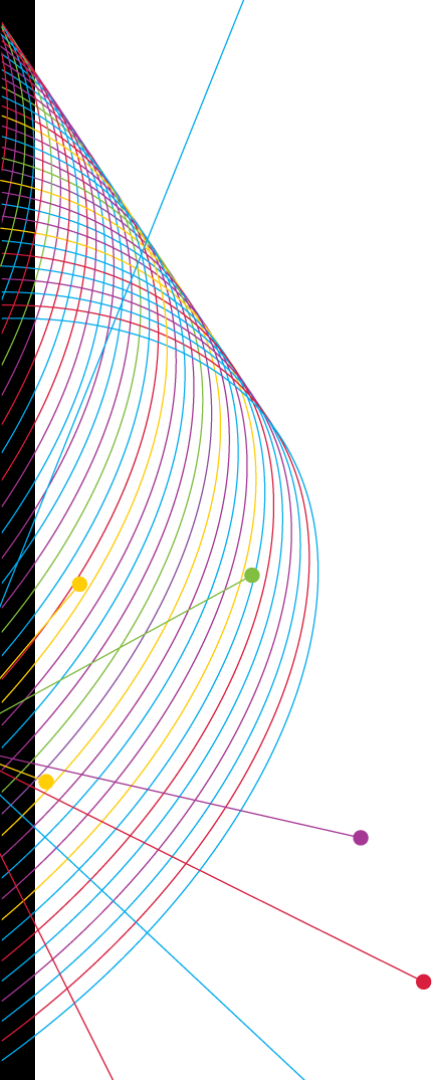
- Example:

	O+	O-
repurchase	0.030759	0.009670
fell	0.020964	0.039305

Penalty Term

- Combining the penalty term, the estimated value, p, could be shrunk to 0.5, which avoid the extreme result:

$$\lambda \log(p_i(1-p_i))$$



PART 3. EMPIRICAL ANALYSIS

MATCH TEXT DATA WITH RETURN

Get the excessive return of day t+1

Excessive Return = True Return − S&P 500 return

All returns are calculated on (Day t+1).

Release before 15:00 ?

- The day after release — N
- The release day — Y

NEWS DATA

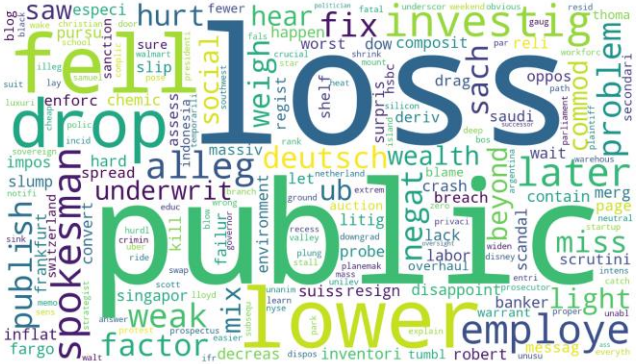
Example of news data collected from Thomson Reuters

	ticker	head	description	txt	datadate	ret
0	XXII	BRIEF-22nd Century ?CEO Henry Sicignano To Dis...	* 22ND CENTURY - CO'S CEO HENRY SICIGNANO III ...	March 19 (Reuters) - 22Nd Century Group Inc: *...	20180319	0.014204
1	XXII	BRIEF-22ND CENTURY GROUP TO DISCONTINUE U.S. S...	* 22ND CENTURY GROUP INC - WILL DISCONTINUE U....	Nov 22 (Reuters) - 22nd Century Group Inc: * 2...	20171122	0.017992
2	XXII	BRIEF-22nd century receives guidance from FDA ...	* 22Nd century receives guidance from FDA on p...	June 22 (Reuters) - 22nd Century Group Inc : *...	20170622	0.022555
3	XXII	BRIEF-22nd Century entered into warrant exerci...	* 22nd Century - entered into warrant exercise...	June 19 (Reuters) - 22nd Century Group Inc: * ...	20170619	-0.024741
4	XXII	BRIEF-22nd Century, Dent Neurosciences Researc...	* 22nd century and Dent Neurosciences Research...	June 8 (Reuters) - 22nd Century Group Inc: * 2...	20170608	-0.000267

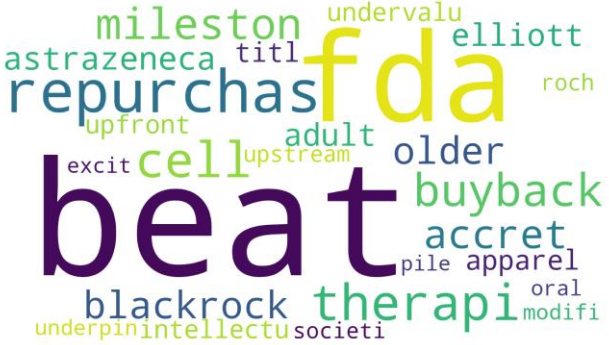
CHOOSING SENTIMENTAL-CHARGED WORDS

Threshold: times = 400 posi_proportion = 0.54 nega_proportion = 0.46

Negative Words Cloud
Number: 296



Positive Words Cloud
Number: 26



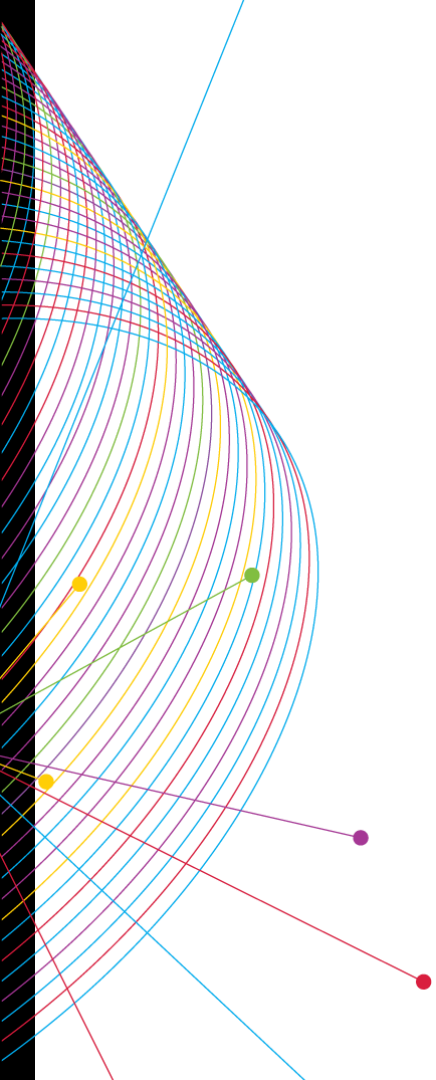
DOCUMENT VECTOR

Construct sentiment word vectors for each news document

News 1:
"Apple just releases a milestone product that beats all the competencies and reduce this year's loss."

News 2:
"Microsoft has to face a drop due to their loss on new products this year. Stock price may drop by 5%"

	Milestone	Beat	Loss	Drop
News 1	1	1	1	0
News 2	0	0	1	2



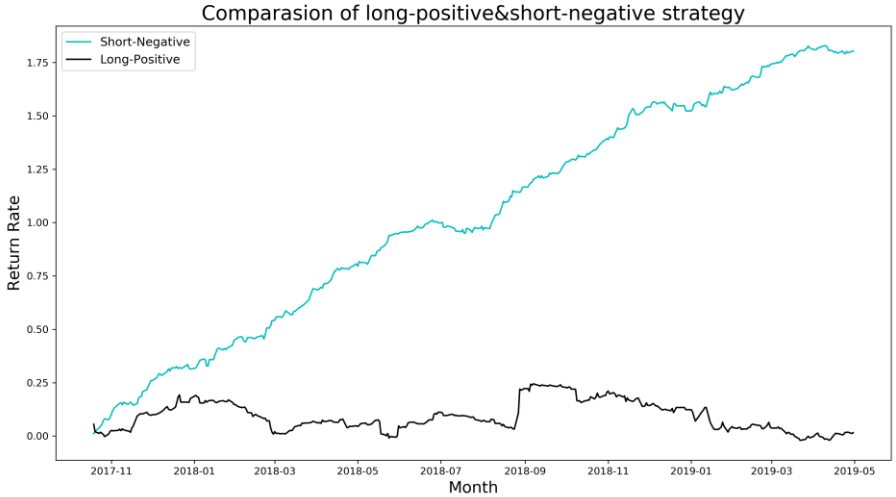
PART 4. PREDICTION AND RESULTS

PREDICTION AND RESULTS

Investor are more sensitive about negative news.

Larger Effect of the Negative News

- **Short-Negative**: **Short** all the related stocks with the predicted **negative** scores
- **Long-Positive**: **Long** all the related stocks with the predicted **positive** scores

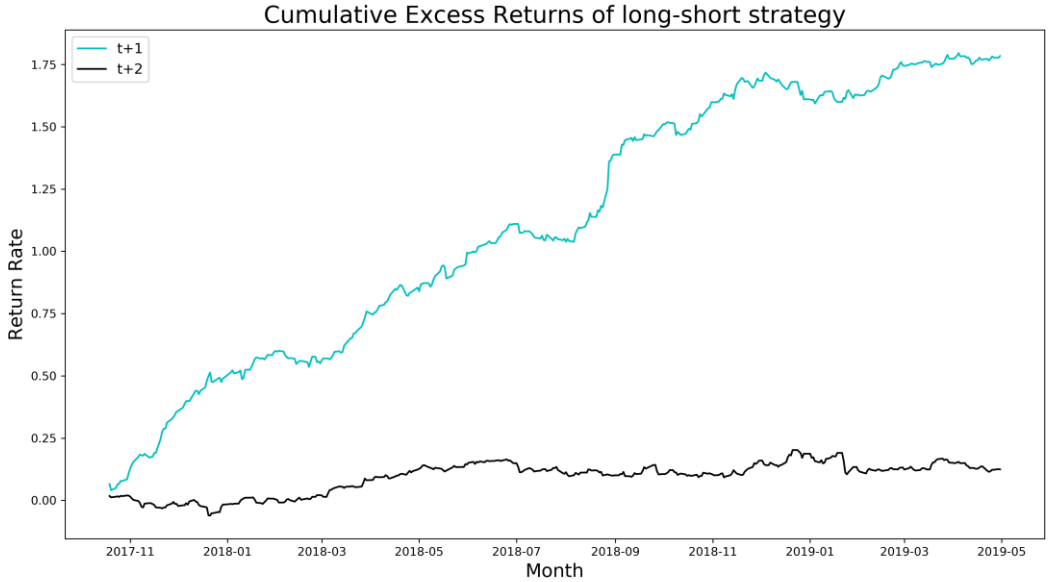


PREDICTION AND RESULTS

Based on the predicted absorption speed of the news, market is fairly efficient.

The timeliness of news

- **T+1**: calculated with **t+1** returns of the related stocks (Correlation: **0.043**)
- **T+2**: calculated with **t+2** returns of the related stocks (Correlation: **0.002**)



FUTURE EXPECTATION

- Figure out a better method to decide the threshold for word subsets selection.
- Limitations of bag-of-words: Meaning (Context, semantic, tone)
- Potential application in the other areas

Thank you!

Q & A