Generation of Supervised Sentiment Metrics for Return and Volatility Prediction from 10-K Filings

Sean Choi (s2101367)
Supervisor: Dr Felipe Costa Sperb

Table

- Background
- Objective
- Contribution
- Methodology
- Experiment Results

Motivation & Background

- From exploratory literature review, we identified an **innovative methodology** by Zheng et al (2020) that utilises stock returns to train sentiments derived from news articles.
- However, one limitation of this approach is that the **sheer volumes of news articles** make it **challenging** to pinpoint those that truly influence **stock returns**.
- Notably, researches suggest that investors primarily rely on **10-K reports** to inform their decisions.

Objectives

Generating **a data pipeline** to collect and orchestrate **a large number of 10-K reports** related to firms in **the technology sector**, in order to use Zheng et al.'s approach to generate **sentiment score** for **(a) the whole report** and **(b) specific to Item 1A: Risk factor**s, both trained on stock returns and on volatility from there companies.

Novelty & Contributions

- **The 10-K filling** has been used to generate sentiment metrics for supporting return and volatility prediction -> No prior study used SEC fillings for it

- **Innovative model**(Zheng et al, 2020) we referred with 10-K fillings - less expensive, white box, scalability, automatised labelling -> No prior study used 10-K filling and its risk factor section unlike the model using news article

- Targeting **volatility**-predicative sentiment signals as well return one unlike Zheng et al (2020)

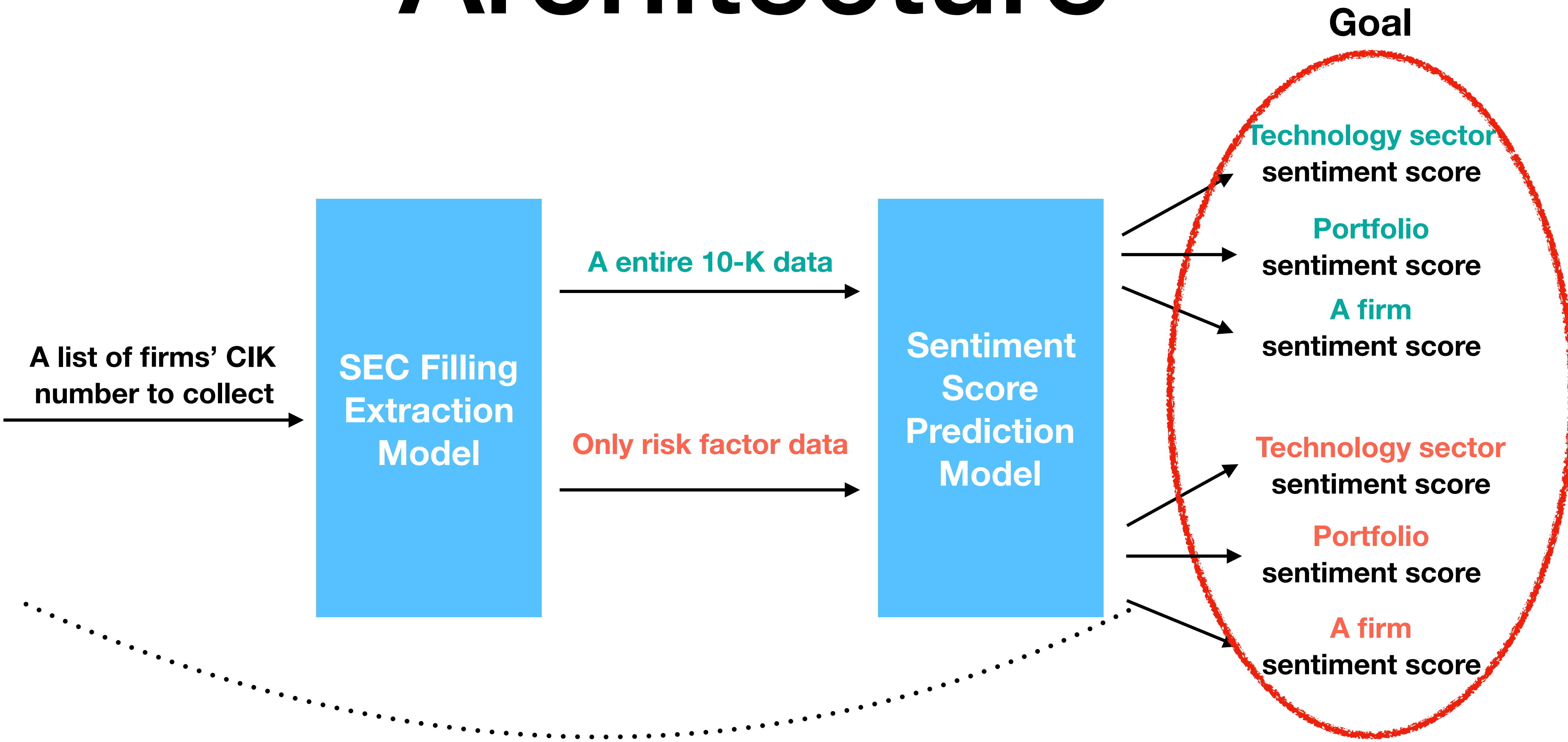
- Generate sentiment metrics on **three key stakeholder** levels: sector, portfolio, and firm -> offering multidimensional sentiment metrics for investors' informed decision making

- Critical evaluation of sentiment metrics through both **quantitative analysis** (i.e. correlation analysis) and **qualitative analysis** (i.e. most influential words to return/volatility)

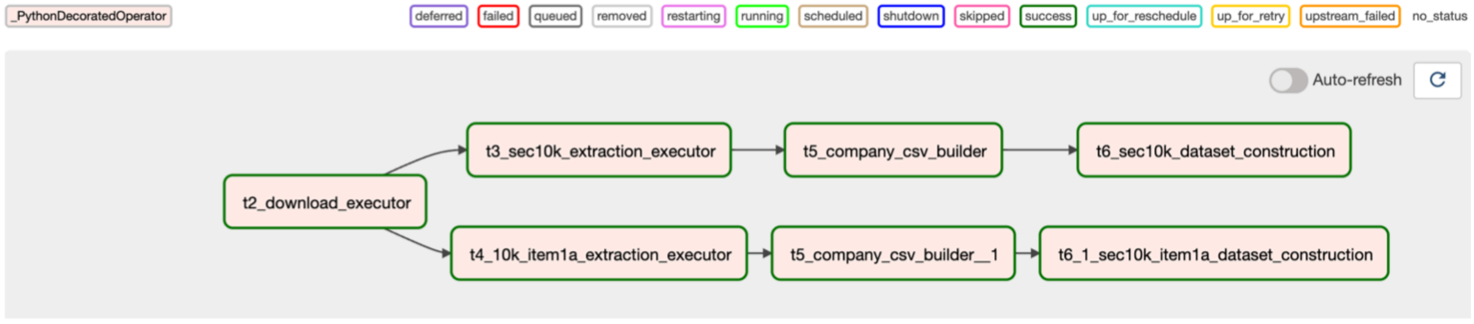
- Developed **an automated data orchestration** where the SEC filling extraction model is seamlessly integrated with the Sentiment Score Prediction Model, updating the **latest** sentiment information immediately when fillings are released.

Methodology

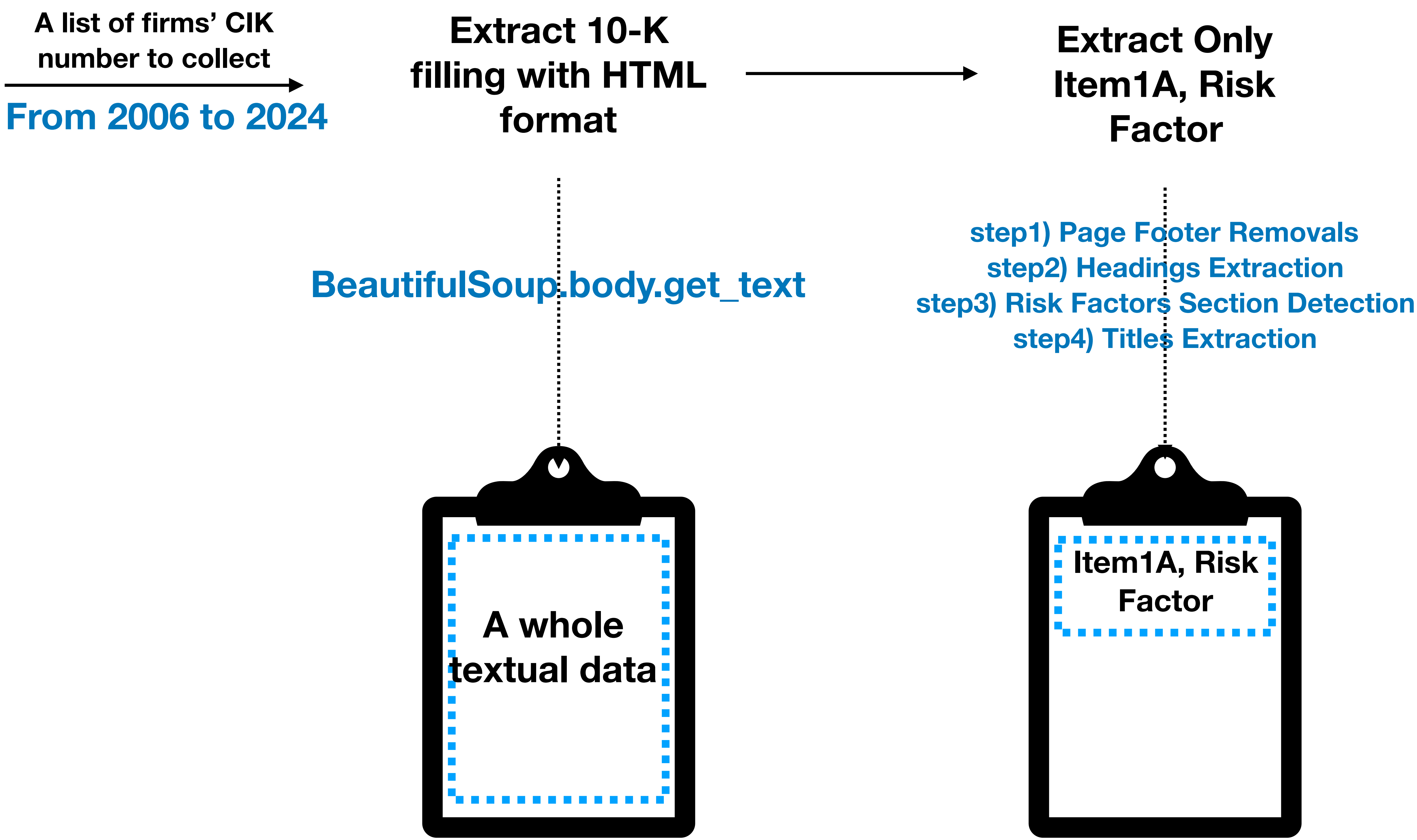
Architecture



Automated data pipeline via Airflow



SEC Extraction Model



Sentiment Score Prediction Model

10-K Textual data →

d_1 e.g) Nvidia's 2006 10-K

w1. w2. w3. w4. w5. ...

w11, w12, w13, w14, w15, ...

Red: + Blue: -

→ **Extracting the sentiment charged word via (4.2) and (4.3)** → $\hat{S}_n$ = [w1,w2,w4,...] → **Estimating vectors of the sentiment topics $\hat{O}$ via (4.5) and (4.8) - return (4.10) and (4.8) - volatility** → $\hat{O}$ → **Estimating the sentiment score via (4.9)**

(4.2): $$f_w = \frac{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} * 1_{\{y_i>0\}}}{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}}}$$

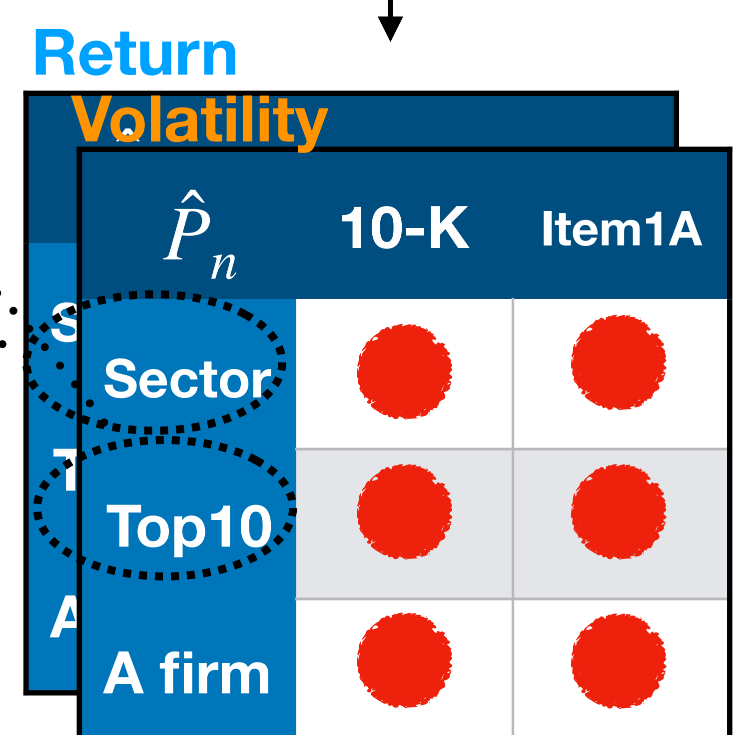
(4.3): $$\hat{S} = \left(\{w : f_w > 0.5 + \alpha^+\} \cup \{w : f_w < 0.5 - \alpha^-\}\right) \cap \left\{w : \sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} > K\right\}$$

(4.5): $$\check{p}_i = \frac{\textbf{rank}(y_i)}{n}$$ **(4.10):** $$\check{p}_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}$$

(4.8): $$\hat{O} = [\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n]\hat{W}^T\left(\hat{W}^T\hat{W}\right)^{-1}, \text{ where } \hat{W} = \begin{bmatrix} \check{p}_1 & \check{p}_2 & \ldots & \check{p}_n \\ 1-\check{p}_1 & 1-\check{p}_2 & \ldots & 1-\check{p}_n \end{bmatrix}$$

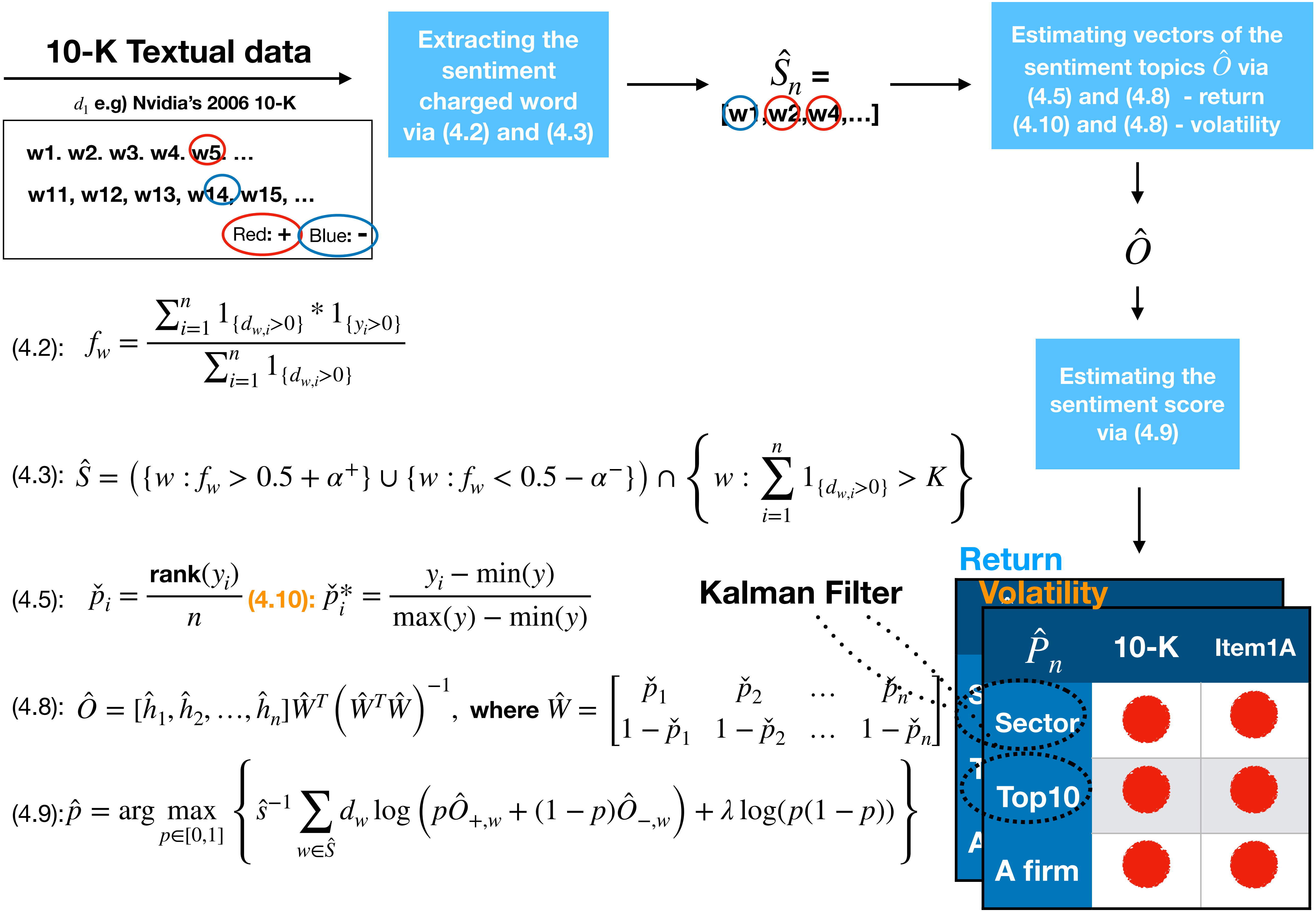
(4.9): $$\hat{p} = \arg\max_{p\in[0,1]} \left\{\hat{s}^{-1}\sum_{w\in\hat{S}} d_w \log\left(p\hat{O}_{+,w} + (1-p)\hat{O}_{-,w}\right) + \lambda\log(p(1-p))\right\}$$

Kalman Filter



*Pink refers to typo in the report

Sentiment Score Prediction Model



(4.2): $$f_w = \frac{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} * 1_{\{y_i>0\}}}{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}}}$$

(4.3): $$\hat{S} = \left(\{w : f_w > 0.5 + \alpha^+\} \cup \{w : f_w < 0.5 - \alpha^-\}\right) \cap \left\{ w : \sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} > K \right\}$$

(4.5): $$\check{p}_i = \frac{\textbf{rank}(y_i)}{n}$$ (4.10): $$\check{p}_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}$$

(4.8): $$\hat{O} = [\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n]\hat{W}^T\left(\hat{W}^T\hat{W}\right)^{-1}, \text{ where } \hat{W} = \begin{bmatrix} \check{p}_1 & \check{p}_2 & \ldots & \check{p}_n \\ 1-\check{p}_1 & 1-\check{p}_2 & \ldots & 1-\check{p}_n \end{bmatrix}$$

(4.9): $$\hat{p} = \arg\max_{p\in[0,1]} \left\{ \hat{s}^{-1} \sum_{w\in\hat{S}} d_w \log\left(p\hat{O}_{+,w} + (1-p)\hat{O}_{-,w}\right) + \lambda \log(p(1-p)) \right\}$$

How to calculate y_i (i.e., Return or Volatility)

Sentiment Proxy

Return Sentiment Proxy: $\check{p}_i = \frac{\text{rank}(y_i)}{n}$ **(4.5)**

Volatility Sentiment Proxy: $\check{p}_i^* = \frac{y_i^* - \min(y)}{\max(y) - \min(y)}$ **(4.10)**

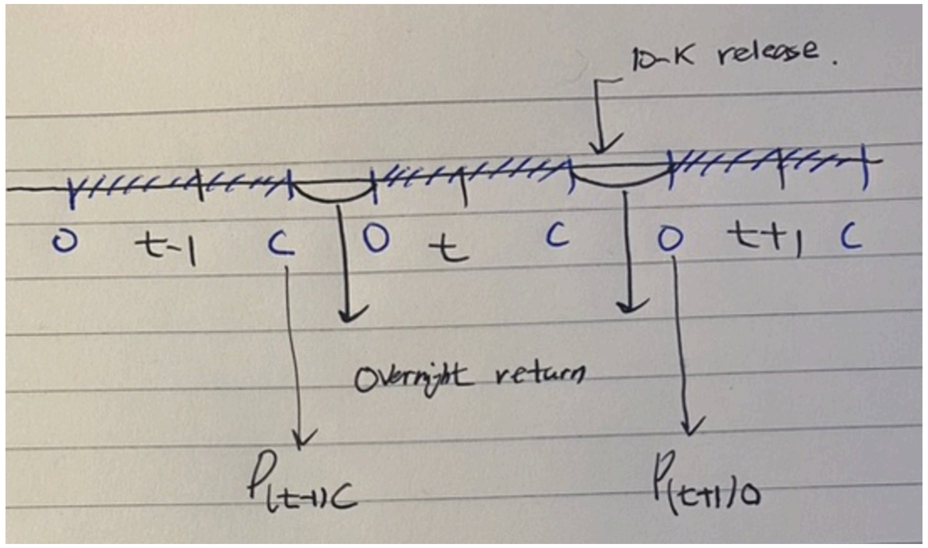


Figure 1

Return

$$y_i = R_t = \log\left(\frac{P_{(t-1)c}}{P_{(t+1)o}}\right) \quad \textbf{(4.4)}$$

Volatility

$$RG_t = \max_\tau \log P_\tau - \min_\tau \log P_\tau, \quad \tau \in [t_o, t_c]\,. \quad \textbf{(4.11)}$$

$$\tilde{V}_t = \frac{RG_t^2}{4\log(2)} \quad \textbf{(4.12)}$$

$$y_i^* = V_t = \frac{1}{3}\left(\tilde{V}_{t-2} + \tilde{V}_{t-1} + \tilde{V}_t\right) \quad \textbf{(4.13)}$$

***Pink refers to typo in the report**

Sentiment Score Prediction Model

10-K Textual data

d_1 e.g) Nvidia's 2006 10-K

w1. w2. w3. w4. w5. ...

w11, w12, w13, w14, w15, ...

Red: + Blue: -

Extracting the sentiment charged word via (4.2) and (4.3)

$\hat{S}_n$ = [w1,w2,w4,...]

Estimating vectors of the sentiment topics $\hat{O}$ via (4.5) and (4.8) - return (4.10) and (4.8) - volatility

$\hat{O}$

Estimating the sentiment score via (4.9)

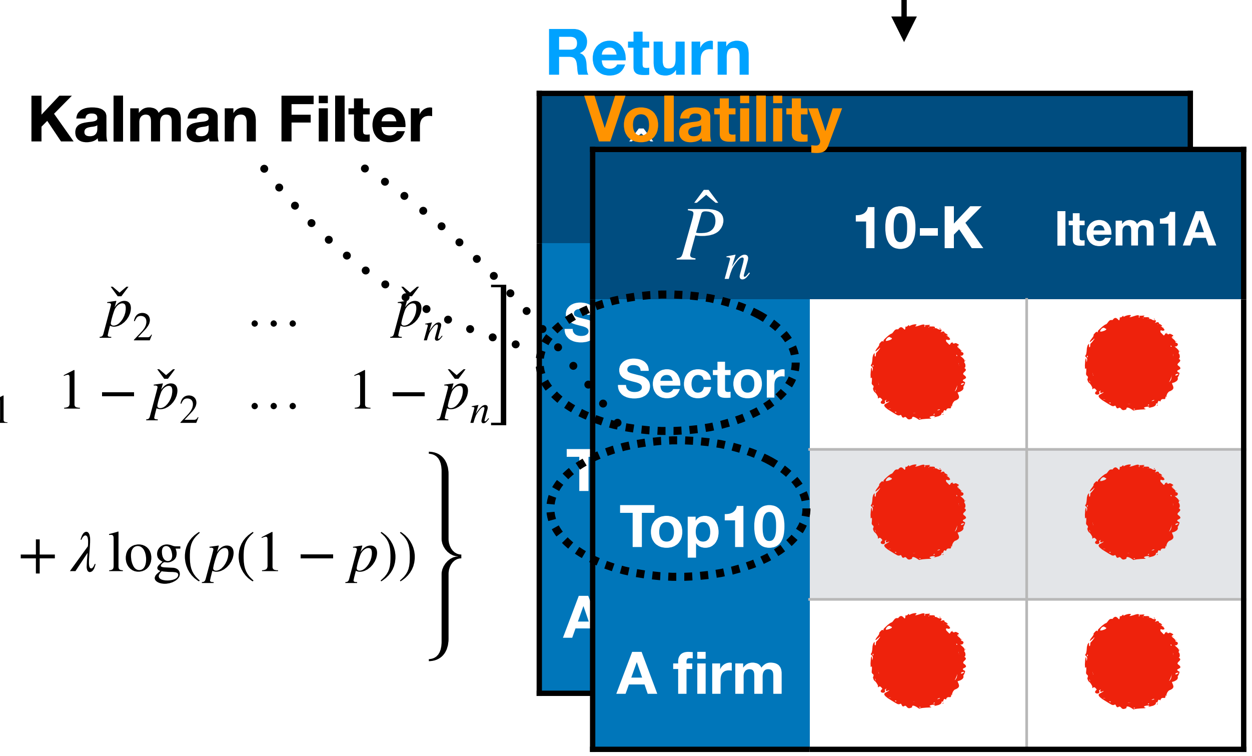
(4.2): $$f_w = \frac{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} * 1_{\{y_i>0\}}}{\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}}}$$

(4.3): $$\hat{S} = \left(\{w : f_w > 0.5 + \alpha^+\} \cup \{w : f_w < 0.5 - \alpha^-\}\right) \cap \left\{ w : \sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} > K \right\}$$

(4.5): $$\check{p}_i = \frac{\textbf{rank}(y_i)}{n}$$ (4.10): $$\check{p}_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}$$

(4.8): $$\hat{O} = [\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n]\hat{W}^T \left(\hat{W}^T \hat{W}\right)^{-1}, \text{ where } \hat{W} = \begin{bmatrix} \check{p}_1 & \check{p}_2 & \ldots & \check{p}_n \\ 1-\check{p}_1 & 1-\check{p}_2 & \ldots & 1-\check{p}_n \end{bmatrix}$$

(4.9): $$\hat{p} = \arg\max_{p\in[0,1]} \left\{ \hat{s}^{-1} \sum_{w\in\hat{S}} d_w \log\left(p\hat{O}_{+,w} + (1-p)\hat{O}_{-,w}\right) + \lambda \log(p(1-p)) \right\}$$



***Pink refers to typo in the report**

Experiment Results

Descriptive Statistics

	Mean	Standard Deviation
Sector Level (i.e., Technology Sector)		
$\tilde{p}^{RET}$ with 10-K	0.48	0.28
$\tilde{p}^{RET}$ with 10-K	0.43	0.24
$\tilde{p}^{VOL}$ with Risk Factor	0.46	0.27
$\tilde{p}^{VOL}$ with Risk Factor	0.42	0.25
Portfolio Level (i.e., Top10 Firms Portfolio)		
$\tilde{p}^{RET}$ with 10-K	0.42	0.30
$\tilde{p}^{RET}$ with 10-K	0.43	0.33
$\tilde{p}^{VOL}$ with Risk Factor	0.42	0.29
$\tilde{p}^{VOL}$ with Risk Factor	0.44	0.31
Company Level (i.e., Nvidia)		
$\hat{p}^{RET}$ with 10-K	0.49	0.32
$\hat{p}^{RET}$ with 10-K	0.39	0.21
$\hat{p}^{VOL}$ with Risk Factor	0.49	0.32
$\hat{p}^{VOL}$ with Risk Factor	0.41	0.18

Sentiment Scores Correlation Analysis

Entire 10-K Filing		Technology Sector				Portfolio of Firms				Nvidia Only			
		p-RET	p-VOL	p-LM	Stock Price	p-RET	p-VOL	p-LM	Stock Price	p-RET	p-VOL	p-LM	Stock Price
Technology Sector	p-RET	1											
	p-VOL	**-0.245	1										
	p-LM	**-0.359	**-0.173	1									
	Stock Price	**0.296	**0.121	**-0.91	1								
Portfolio of Firms	p-RET					1							
	p-VOL					**0.905	1						
	p-LM					**-0.922	**-0.903	1					
	Stock Price					**0.956	**0.824	**-0.841	1				
Nvidia Only	p-RET									1			
	p-VOL									*0.612	1		
	p-LM									**-0.832	*-0.482	1	
	Stock Price									0.234	0.588	0.182	1

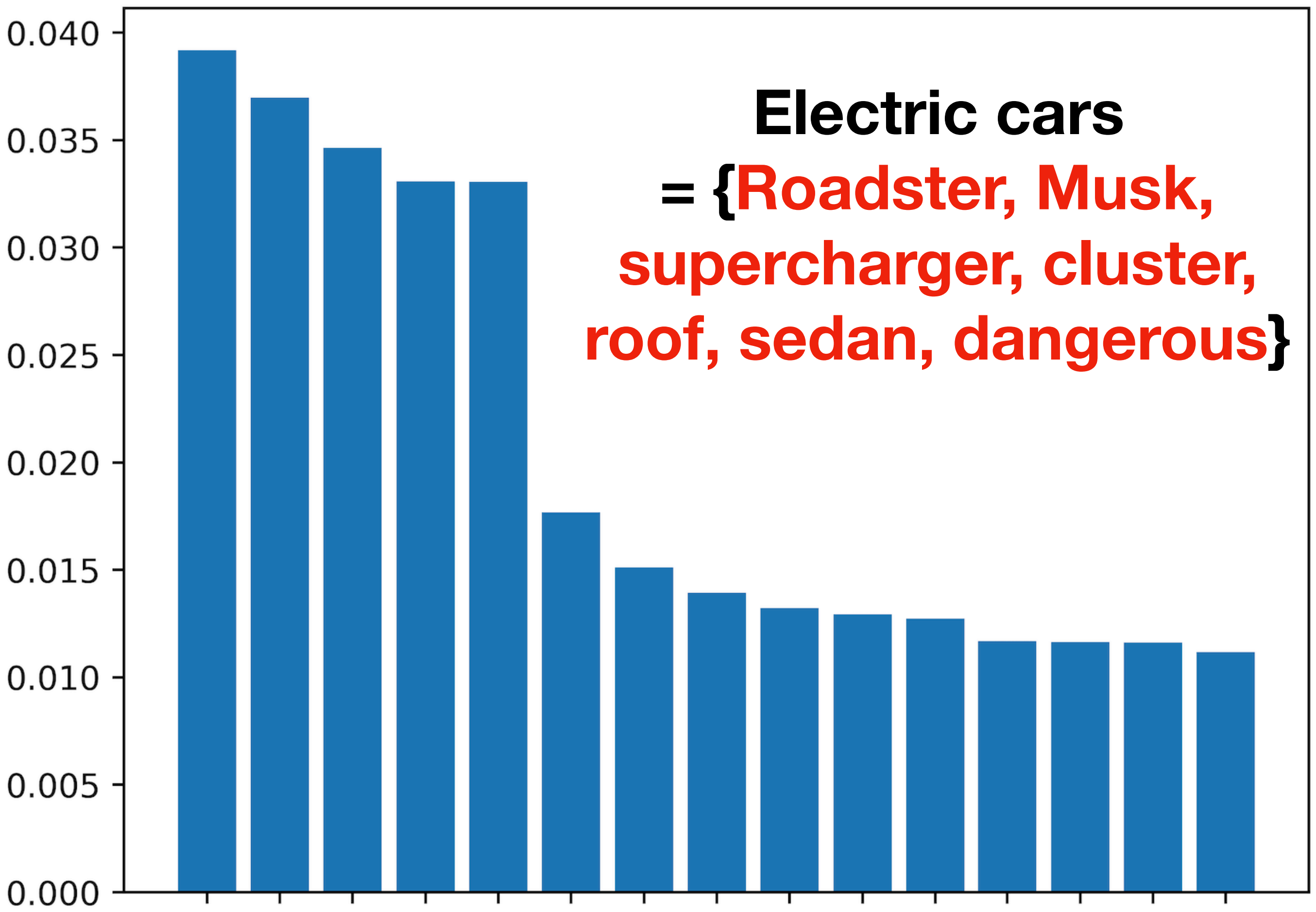
Note: * p-value < 0.05, ** p-value < 0.005

Item 1A (Risk Factors)		Technology Sector				Portfolio of Firms				Nvidia Only			
		p-RET	p-VOL	p-LM	Stock Price	p-RET	p-VOL	p-LM	Stock Price	p-RET	p-VOL	p-LM	Stock Price
Technology Sector	p-RET	1											
	p-VOL	**-0.167	1										
	p-LM	0.008	**0.182	1									
	Stock Price	**-0.376	*0.056	**-0.764	1								
Portfolio of Firms	p-RET					1							
	p-VOL					0.145	1						
	p-LM					**-0.806	**-0.349	1					
	Stock Price					**0.893	*0.185	**-0.937	1				
Nvidia Only	p-RET									1			
	p-VOL									*0.490	1		
	p-LM									0.057	0.004	1	
	Stock Price									0.228	**0.712	*-0.593	1

Note: * p-value < 0.05, ** p-value < 0.005

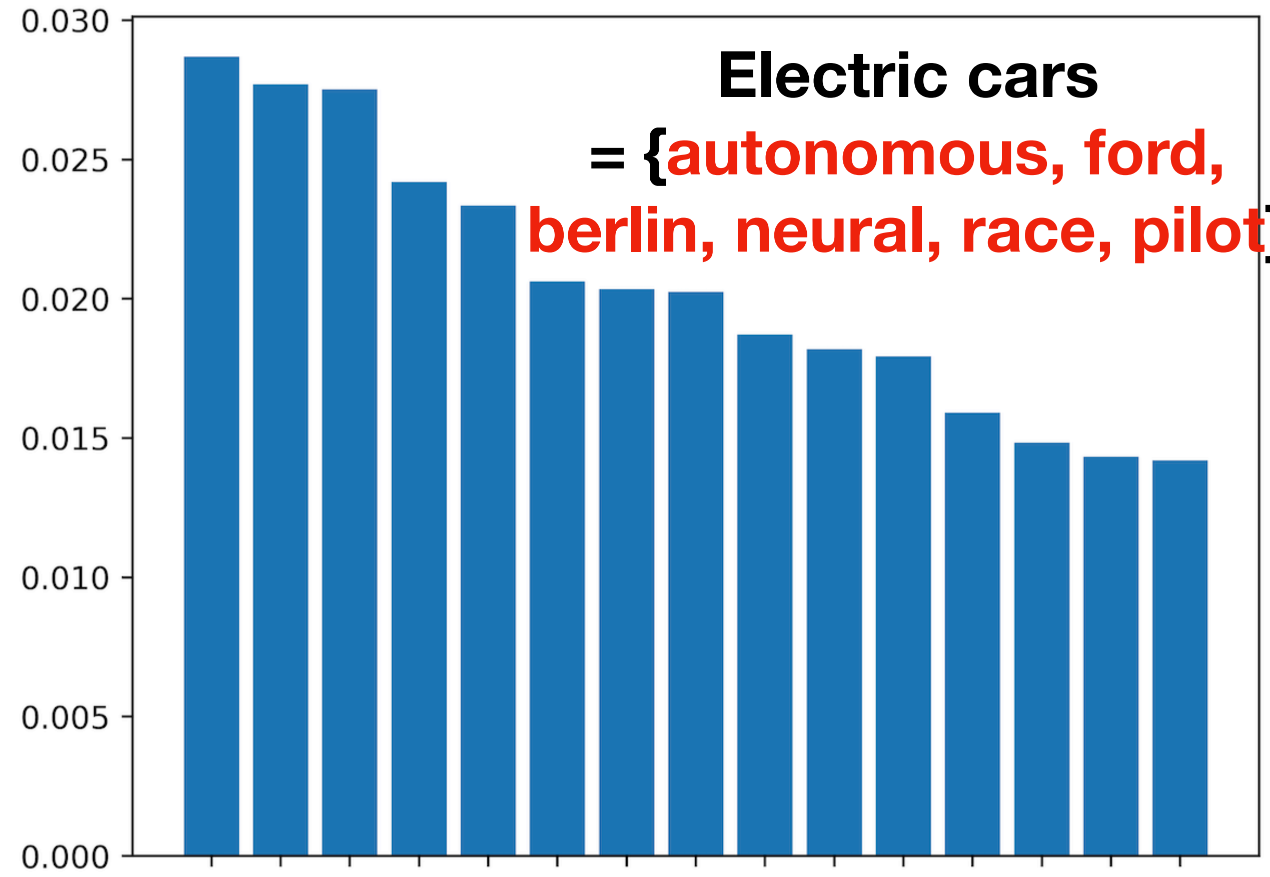
Exploratory Dictionary Orientation Analysis: Portfolio Level

The influential words in $\tilde{p}^{RET}$ with 10-K , positively



roadster, misconduct, musk, supercharger, agricultural, detriment, cluster, owing, observe, rigor, roof, sedan, visible, ramping, dangerous

The influential words in $\tilde{p}^{VOL}$ with 10-K , positively



autonomous, ford, draft, berlin, thin, neural, audible, accomplished, log, understatement, identifier, segregation, subjectivity, race, pilot

*Pink refers to typo in the report

Limitations

- The mode can not capture meaningful phrase word.
- The model does not consider the allocation proportion of the QQQ portfolio.
- The model can not adapt to the latest firms' return or volatility until the fillings are released at the publication date

References

Zheng Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting returns with text data. University of Chicago, Becker Friedman Institute for Economics Working Paper, (2019-69), September 2020. Yale ICF Working Paper No. 2019-10, Chicago Booth Research Paper No. 20-37.