Honours Project

Sean Choi (s2101367)
Supervisor: Dr Felipe Costa Sperb

Hi.
Happy to introduce my honours project.

Table

- Background
- Objective
- Model Architecture
- Experiment Results
- Timeline

Table
My 10 min presentation will be consist of background explanation, goal, Model Architecture, Experiment Results and timeline plan.
As I have introduced background, and goal on the previous presentation, today I will focus on model architectures and experiment results.

Motivation & Background

- The 10-K is a comprehensive official document that offers a thorough overview of a company's business.
- Section 1A on the risk factors contains the most significant elements that may make the company speculative or risky.
- Several studies argue the risk factor disclosure provides valuable information.

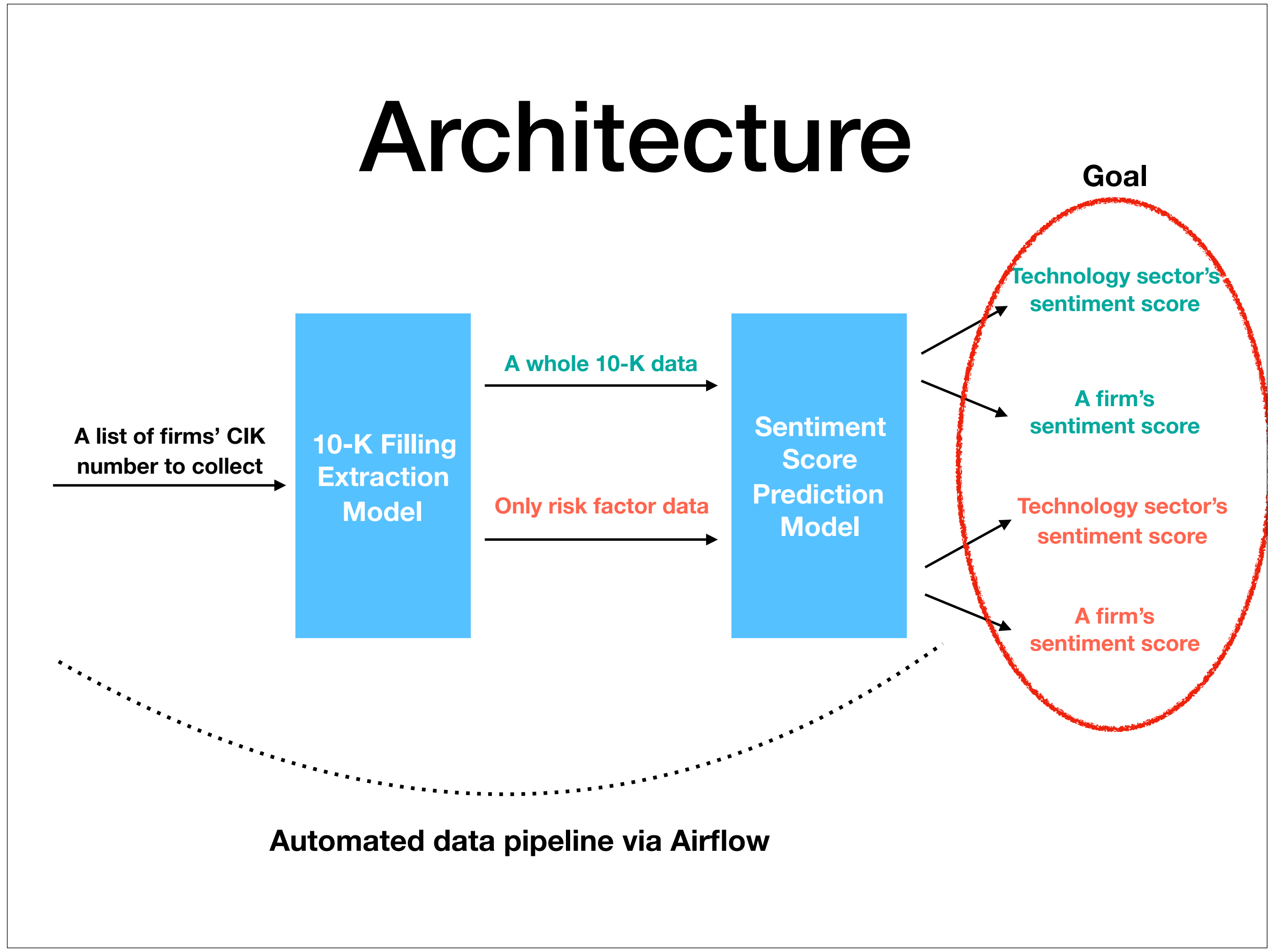
My initial motivation comes from the nature of the 10-K report.
The 10-K report is a comprehensive official document illustrating a company's business. In my research, I will use both a whole 10-K filling and the section 1A on the risk factors. Section 1A contains the most significant elements that may make the company speculative or risky.

~~The 10-K report is a comprehensive official document submitted to SEC, which is US government agency. It offers a thorough overview of a company's business. Among them, the section 1A on the risk factors contains the most significant elements that may make the company speculative or risky. Several studies support the value of the risk factor disclosure providing richer and valuable information.~~

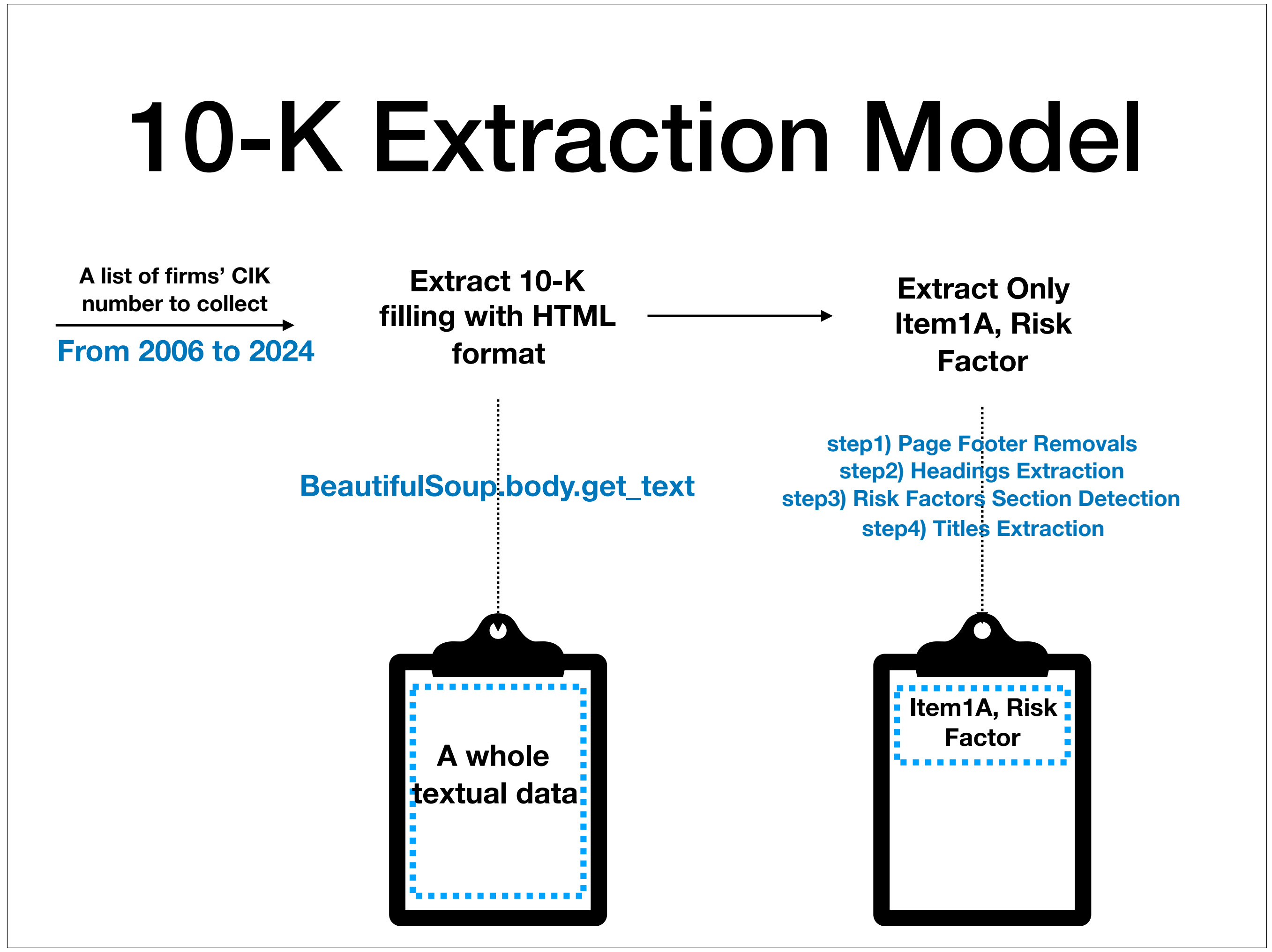
Objectives

Generating a valuable **metric** (i.e. **sentiment score**) from **10-K** reports for predicting returns and volatility in the **technology industry**.

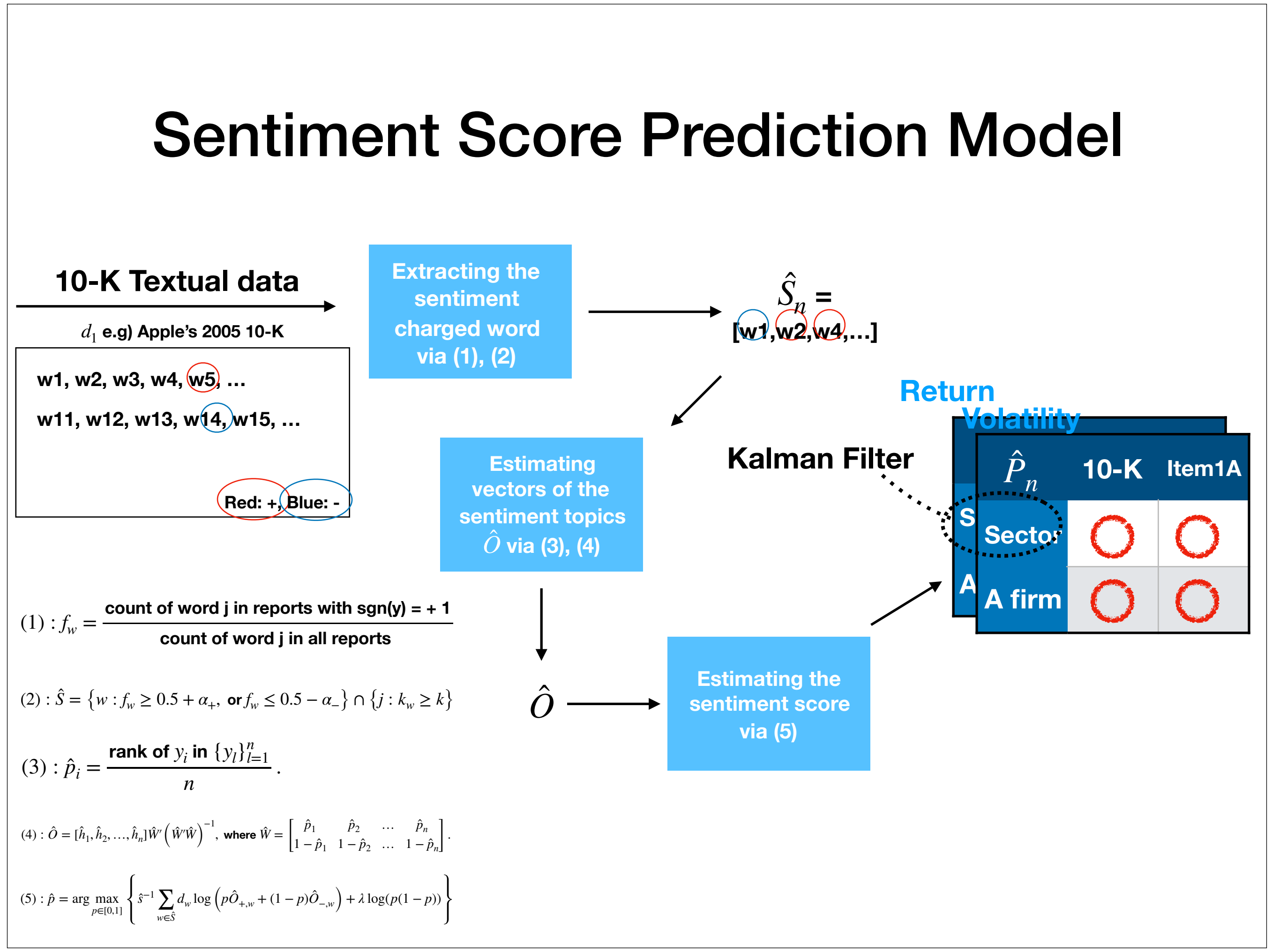
With the valuable 10-K report focusing both a 10-K filling itself and the section 1A of the risk factor, I would like to generate a valuable metric for predicting returns and volatility in the technology industry. In this project, I will focus on generating sentiment scores from 10-K report.



Sentiment Score Prediction Model consists of the two stages: data extraction and Sentiment Score Prediction. And, all processed will be connected to Yong's Airflow server, so that users such as analyst or investors can attain the latest updated sentiment scores of the 10-K fillings. To make the model more useful for users, the model can show both macroscopic and microscopic sentiment scores. In other words, the model predicts both the sentiment scores of 100 firms representing technology sector, and the sentiment scores of a single firm within the tech sector. Moreover, we used a whole textual data of the 10-K fillings as well as the section 1A risk factor part to overcome insufficient sample dataset. We assume that using only risk-factor section would not produce informative information about the sentiment for a firm. Users potentially can analyse the sentiment of a firm with various perspectives with the aid of our model.



When extracting 10-K reports and its subparts such as Item 1A risk factor, we can relatively easily collect it due to the uniform structures of the 10-K fillings. Form 10-K fillings can be easily collected by using a Central Index Key(CIK)in EDGAR. A CIK is a number given to an individual or company by the SEC, used to identify the ownership of a filling. Initially, we collected the CIKs list of firms listed in QQQ to facilitate data retrieval. After retrieving a whole textual data of 10-K fillings, we extracts Item 1A, Risk factor section through four steps: Firstly, we remove page footers. Secondly, we extract headings within the filling. Thirdly, we detect risk factors section with the heading extracted. Finally, we extract titles in the docs.



This is an overview of generating sentiment scores. Previously, I have introduced the methodology. I will briefly explain it in today's presentation.
After extracting 10-K fillings and extract risk factors section from them.

We will filter the sentiment charged-word via equation (1), and equation (2). Equation (1) defined the frequency of each word w in 10-K reports with positive returns compared to its overall frequency in all reports. Equation (2) is about identifying sentiment-charged words. It involves setting thresholds to determine whether a word is positively or negatively charged in sentiment. We filter out the set of neutral words through this equation. If f_w equals to 0.5, it means the word w has a neutral tone, but we remove the set of natural words to remove the noisy.

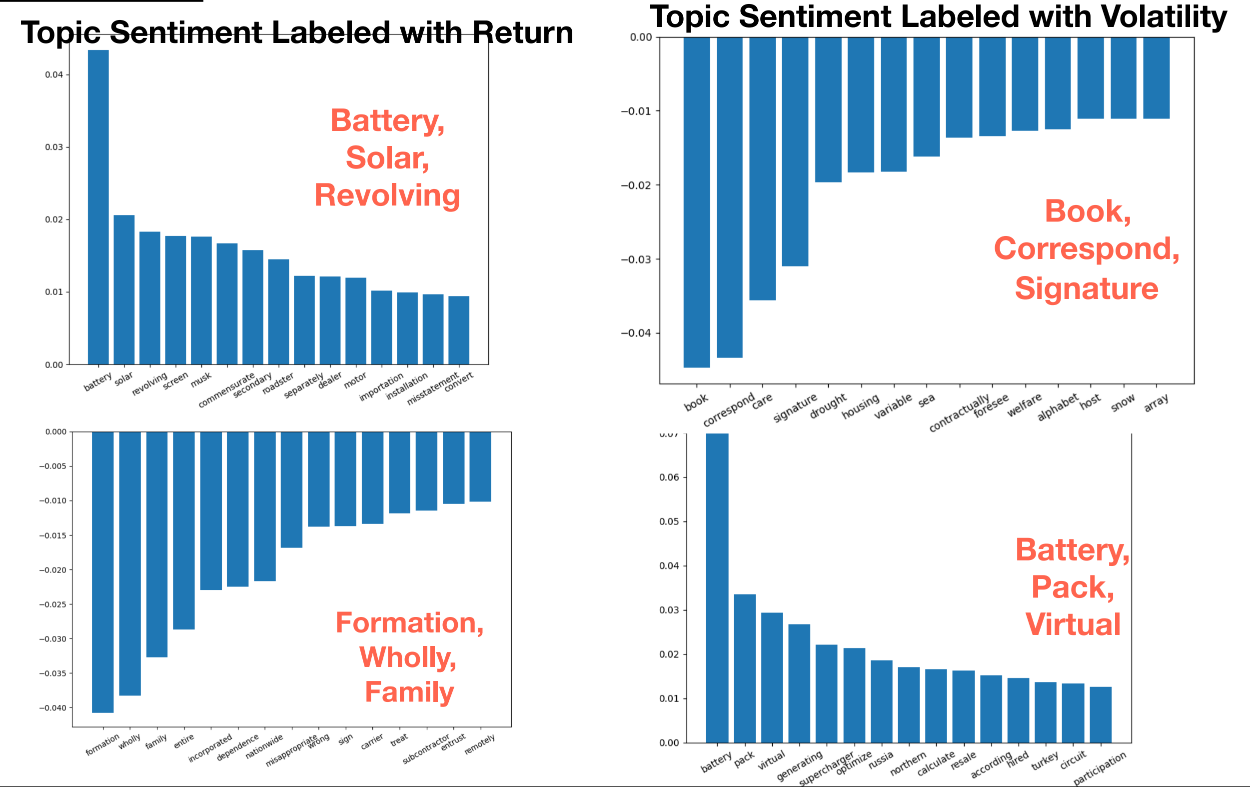
After attaining the sentiment charged-words for a single report, we will estimate vectors of the sentiment scores via equation (3) and (4).
The equation (3) represents the standardised rank of the return associated with the i th report relative to the total number of 10-K reports. ~~For instance. Let's say we have collected Apple's 10-K fillings from 2005 to 2023, and this is one single 2005 report. The equation (3) is then the rank of Apple's 2005 return compared to all returns from 2005 to 2023.~~
The hat of h in the equation (4) is the proportion of sentiment-charged words in the report relative to the total number of words. After (3) and (4) process, we can estimate the sentiment score of a 10-K risk factor. We will generate the sentiment score, which is the p hat by using maximum likelihood optimisation technique. This is the model fitting on return.

Furthermore, we can find the sentiment score fitting on volatility. To do that, we need to tweak the equation (1), (2), and (3) to adapt for volatility fitting.~~ This is because return in nature has binary features; loss and profit, whereas volatility doesn't have. For simple calculations, we want to set binary futures on volatility like high volatility and low volatility based on the 80% quantiles. Above the 80% quantile of the distribution is high volatility, whereas The below is low volatility.~~

Again, our goal of the study is to generate various sentiment scores. Through the model, we can generate eight types of sentiment score metrics. Four metrics labeled on return contains technology sector's sentiment scores by using either 10-K report itself or only Item1A Risk Factor section, and a firm's sentiment scores by using either 10-K report itself or only Item1A Risk Factor. Likewise, we can have four metrics in the context of volatility.

Experiment Results

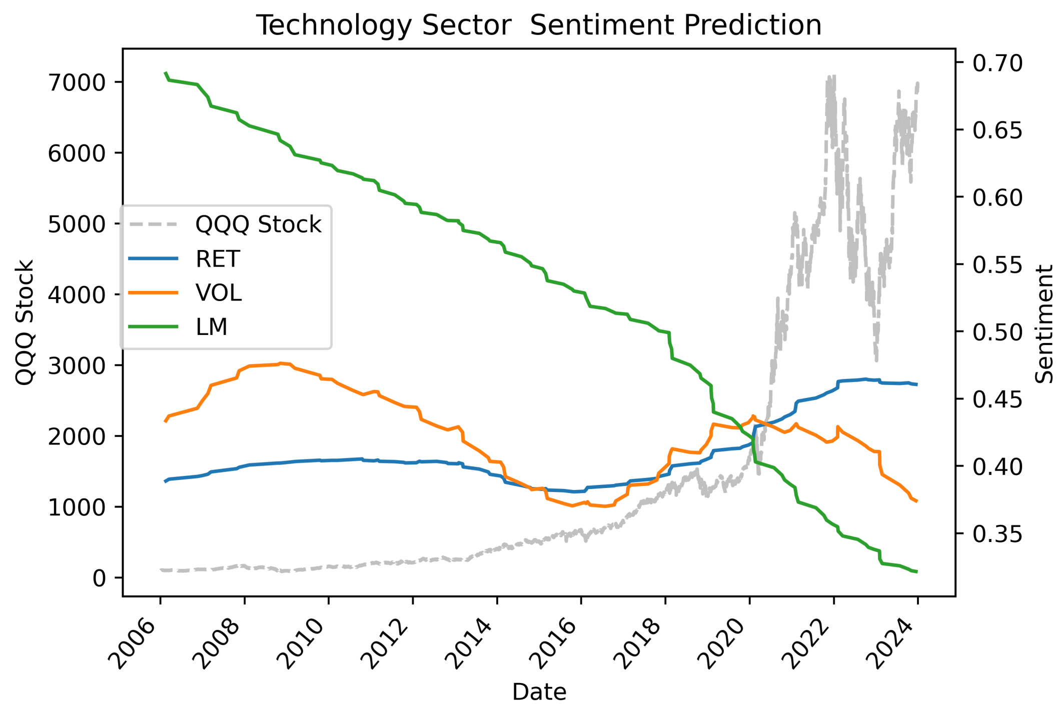
	10-K	Item1A
Sector	○	●
A firm	○	○



This is one of experiment results. We have used the risk factor section of 10-K fillings of firms representing technology sector as a data by labelling either return or volatility. The top left graph shows top 15 words having positive tone when labelled on return. We can see the three words; Battery, Solar, Revolving are the top 3 words having positive tone of the technology sector. The top right graphs shows sentiment words related to low volatility on the technology sector. We can see that Book, Correspond, Signature are related to the lower volatility of the technology sector.

	10-K	Item1A
Sector	○	●
A firm	○	○

Experiment Results



~~This graph shows the predicted sentiment score labeled on either return or volatility. And the grey line refers to the history of QQQ ETF stock. From this graph, we can see the trend of sentiment of the technology sector. Again, high sentiment score on return at the year means the sector has a positive tone on that year, whereas high sentiment score on volatility can show high market uncertainty on that year. The return sentiment score in this experiment tend to align with the QQQ market movement positively correlated. However, it does not represent well the market volatility beginning from 2020.~~

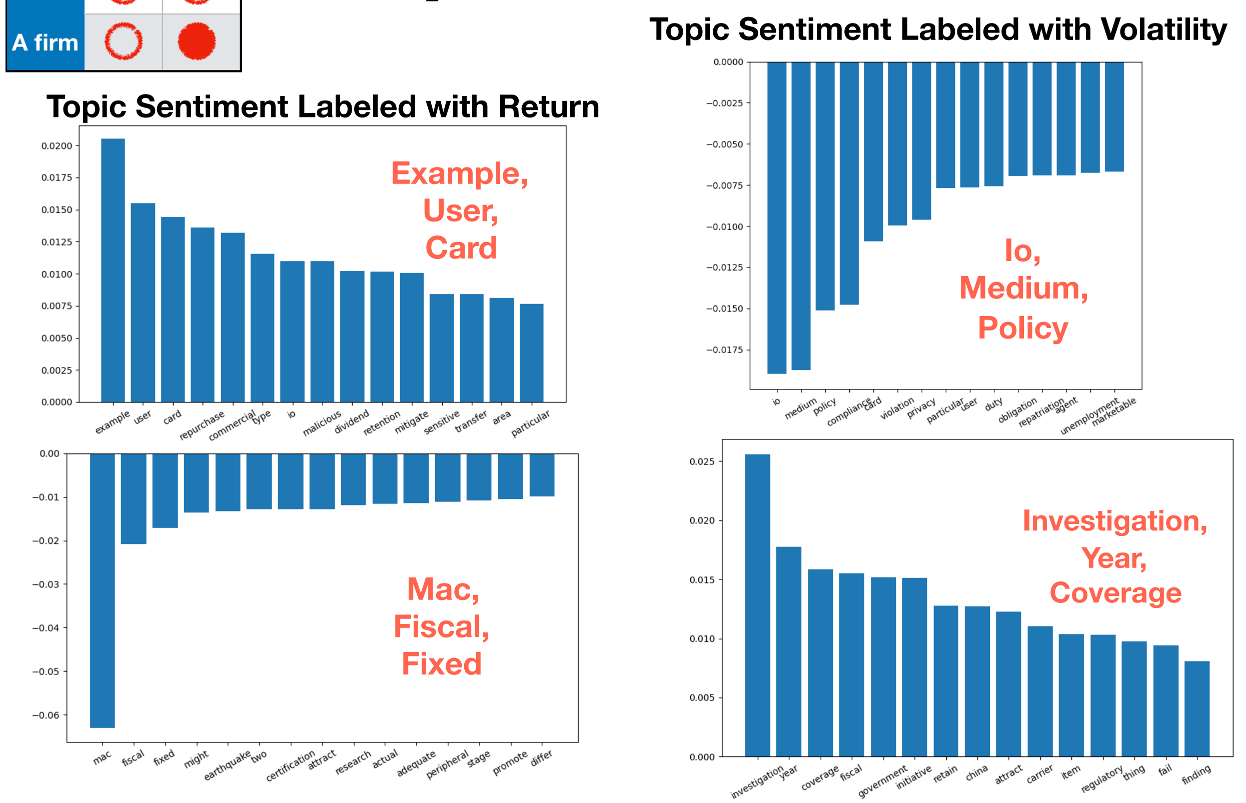
Correlation Between Sentiment and Stock Price: The sentiment score related to returns (RET) seems to have a positive correlation with the QQQ stock price until around 2020, indicating that as the stock price increased or decreased, the sentiment on returns mirrored this movement. However, after 2020, the sentiment does not seem to track with the QQQ stock price, particularly for market volatility (VOL), which may suggest a divergence between actual market conditions and the sentiment or perhaps an increase in market uncertainty that is not reflected in stock prices.

Observations Post-2020: The graph shows a significant change after 2020. The QQQ stock price has a sharp upward trend from 2020 to 2022 and a sharp downward from 2022 and 2023, while the sentiment on returns and volatility does not show a sharp fluctuations that do appear to align with the QQQ stock trend. This could indicate a period of market instability or a shift in the factors that influence market sentiment.

Experiment Results

Apple Inc.(AAPL)

	10-K	Item1A
Sector	○	○
A firm	○	●

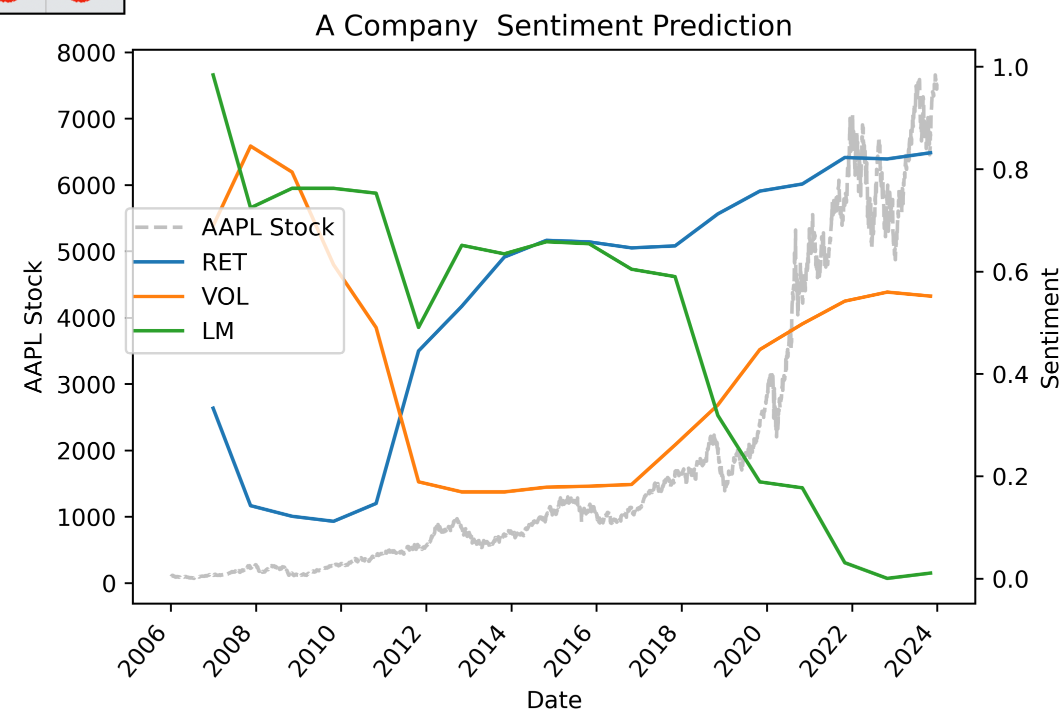


This is example of a single firm's result. In this example, I will introduce Apple's predicted sentiment score.
We can see at the top left the three words; Example, User, Card are the top 3 words having positive tone of the technology sector. The below right graphs shows sentiment words related to high volatility on the technology sector. We can see that Investigation, Year, Coverage are related to the high volatility of the technology sector.

Apple Inc.(AAPL)

	10-K	Item1A
Sector	○	○
A firm	○	●

Experiment Results



There seems to be a correlation between the RET sentiment and the AAPL stock price, especially noticeable from 2016 to 2024, where both the sentiment and stock price trend upwards.

The VOL sentiment support this correlation as it represent market fluctuations. The sharp upward post-2019 represents apple's stock fluctuated.

Limitations

- Phrase word uncaptured
- Insufficient sample data

1. Can not capture meaningful phrase word as the model essentially use a bag-of-words to extract sentiment-charged words. For instance, the model will extract the word ‘chef executive officers’ in a filling separately, and then extract each words with respect to the dependent variables. But, this separation loses an original meaning of a word.

2. Insufficient sample data: either some firms do not have all the 10-K reports released during the periods, or the 10-K fillings tend to be repetitive, although they includes comprehensive and informative information about the firm.

Timeline

- ~~Build the code to generate sentiment scores (Winter vacation)~~
- ~~Build the 10-K report pipeline (Winter vacation)~~
- ~~Implement Case 1 with risk factor (Winter vacation)~~
- ~~Implement Case 2 with risk factor (Winter vacation - W2)~~
- ~~Implement Case 3 with risk factor (W2 - W6)~~
- ~~Implement Case 1,2,3 with a whole 10-K filling (Reading Week)~~
- Connect to Airflow Server (Reading Week)
- Evaluate sentiment scores (W6 - W9) by comparing FinBert
- Wrap up report (W9 - W11)
- Write a report while coding (Always)

This is our planned timeline. Until now, we have used only risk factor data set. To overcome insufficient data set issues, we will add up a whole 10-K filling on our model during the Reading week as well as connecting to Airflow server to automate the sentiment score prediction process. Furthermore, we will evaluate the value of our model by comparing to FinBert, a traditional Finance-relevant pre-trained model.

Thank you.