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

Sanggyu Sean Choi

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2024

Abstract

This paper introduces an automated system that generates sentiment metrics to support the prediction of stock returns and volatility. It focuses on three key stakeholder levels: sector, portfolio, and firm. Our system consists of two main components: the SEC Filing Extraction Model and the Supervised Lexicon Learning Model. The SEC Filing Extraction Model is responsible for preprocessing SEC filings, facilitating seamless integration with the subsequent Supervised Lexicon Learning Model. The lexicon model operates through a four-stage process: (i) identification of sentiment-charged words via predictive filtering, (ii) assignment of prediction weights to these tokens using topic modelling techniques, (iii) estimation of the most probable sentiment score by aggregating the weighted tokens through penalised likelihood, and (iv) application of the Kalman Filter for sector or portfolio sentiment trend analysis. In our empirical study, we study one of the most comprehensive and essential documents about a public firm - 10-K filling, and its Item1A risk factor section. At the sector level, our 10-K-centred model outperforms our risk-factor-centred model in extracting return/volatility-predictive signals in the context. At the portfolio level, both models excel in identifying return/volatility-predictive signals within the context. We recommend, at the company level, the risk model for trend and correlation analysis while advising both models for word analysis.

Keywords Sentiment Analysis, Fundamental Analysis, Data Orchestration, Machine Learning, Return, Volatility, 10-K fillings

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sanggyu Sean Choi*)

Acknowledgements

I thank my supervisor, Dr. Luis Felipe Costa Sperb, for his invaluable support and guidance throughout this project. I sincerely appreciate it. Also, I thank his PhD student, Hao Zhou, for his support.

Many thanks and love to my family(Hyungmook Choi, Jungsun Lee, Jiwon Choi) and my grandparents(Jongmook Lee, Chunja Kwon) for your love.

Foremost, Thank you, God, for your immeasurable love.

Proverbs 9:10

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Now is the age of Artificial Intelligence and Big data. With the advance of computational powers, a large amount of various data, such as text, video, and audio have been used for scientific analysis. Among a myriad of data forms, textual data has gotten the fastest attention in the social science academic field. Textual data's numerical representation for statistical analysis in nature is extremely high in dimensions that empirical study seeking its textual richness should face its dimensionality challenges. Machine learning will be employed to extract richer meaning from textual data for predictive analysis in a high-dimensional data environment [59].

In finance, textual data is commonly employed for predicting market movements[59, 93, 100, 79, 116, 43]. In stock prediction, textual analysis of market sentiment has shown notable success. News data was employed to analyse sentiments in the prediction of short-term stock price movements [100]. Similarly, social media textual data was utilised by integrating social media sentiment and AI [116]. Also, annual report data was used for stock market forecasting [43]. Likewise, we could find myriad types of textual data were used in predicting market movement. Then, what type of textual data can be informative for such a purpose? If you want to invest in a public company in the United States, where can you begin your investment journey?

There are myriad ways to begin your investment in a public company. However, what if you do not know about a firm at all in which you would invest or what if you do just partially know the firm? Then you should first know the firm correctly. If so, where we can find reliable and trustworthy information about a firm? You can find a rich deposit of reliable knowledge on Form 10-K filing. The Form 10-K filing has been mandated by the Securities and Exchange Commission (SEC) since 1934. It has its origins in the the Section 13 or 15(d) of the Securities Exchange Act of 1934. The 10-K is a comprehensive official document offering a through overview of a company's business, its' potential challenges, and its financial performance through the fiscal year. A company's leadership, in the 10-K, provides their perspective on the business outcomes and the factors influencing them [113]. Furthermore, several studies found

that 10-K filings offered predictive power to stock price prediction[9, 122, 61, 12, 57]. [9] found a shred of evidence that the market reacted to 10-K filings in a statistically significant way. [61, 122] showed there was a correlation between the complexity of 10-K filing and stock price volatility. [12, 57] found a positive correlation between 10-K filing and stock price through cutting-edge AI methodologies. Hence, investors should pay attention to 10-K filings.

Since the beginning of 2005, a new section has been required to be included in all firms' annual filings by the SEC. The section called "Section 1A of the Annual Report on Form 10-K" discusses "the most significant factors that make the company speculative or risky" [1]. Prior to this alteration, companies were only obligated to provide this information in their registration when issuing their equity or debt securities. Also, some companies voluntarily offer risk disclosures in the section called "Management's Discussion and Analysis of Financial Condition and Results of Operations(MD&A)"

Opponents of the new disclosure requirements argue that risk factor disclosures are unlikely to offer valuable information. First, risk disclosure can be biased. Managers might resist disclosing negative information about their business or career incentives [117, 94, 35, 65, 66]. Second, managers' overconfidence could make them perceive less risk or overconfident managers could have the illusion that they can effectively manage the risks confronting their firms [18]. Third, managers tended to disclose all possible risks and uncertainties without making precise predictions or providing detailed financial assessments. This practice stems from the fact that companies were not required to predict the possibility that a disclosed risk would actually materialise. Furthermore, there was no obligation for firms to specify the financial influence that a disclosed risk might have on their present or future financial statements [91]. Since 2010, the SEC has warned companies to "avoid generic risk factor disclosure that could apply to any company" [95], and has continuously pushed the precise risk factor disclosures through the comment letter process [76]. Recently, the SEC has been demanding the explicit and specific inclusion of both cyber security risk factor disclosure and climate change risk factor disclosure [63, 87].

Notably, numerous studies reach an opposite conclusion by providing evidence that the risk factor disclosures, in fact, provide valuable information [15, 54, 104, 50, 67, 91, 18]. The disclosures reflect the genuine risks confronting their firms. Also, it might help investors assess the volatility of a firm's cash flows, and tax-related risk factor disclosures offer details about the level of a firm's future cash flows, helping investors incorporate this information into current stock prices [15]. The newly created risk factor disclosures also show a correlation with conventional asset pricing risk factors, indicating that the disclosed factors are valuable for assessing overall risk [54]. cyber security risk factor disclosures increase the risk of a company's stock price declining in the future [104]. Furthermore, it has been revealed that the length of the disclosure is associated with market reaction. lengthier risk factor disclosures have a negative correlation with market reactions [15]. More detailed disclosures tend to generate more profound market reactions [50]. Alterations in the length of risk disclosures also can influence an investor's risk perceptions [67]. From these observations, although disclosures are occasionally seen as generic [91] or susceptible to bias [18], they still provide risk-related information that can assist investors and affect stock values.

1.2 Project Objective

Given the informational value of 10-K filings, as introduced in the motivation section, our project aims to achieve the following main goal:

- Developing an automated pipeline to generate sentiment scores from 10-K reports of firms in the technology industry for market movement prediction of a firm, i.e., return and volatility.

The following are the sub-steps for the main goal:

- Extraction of 10-K fillings, followed by extraction of risk factors from the extracted 10-K fillings. The list of 10-K fillings was based on the Invesco QQQ Trust Series 1 (QQQ).

Our pipeline will collect an annual disclosure from the SEC's Electronic Data Gathering, Analysis, and Retrieval(EDGAR) system and extract the risk factor section of each report. It also will be able to automatically collect the latest 10-K filings for prompt generation of sentiment information.

- Generate various sentiment scores from the extracted 10-K reports.

In this research, our study aims to generate sentiment metrics for market movement prediction of a firm, i.e., return and volatility, from both the entire 10-K report and risk factor disclosures. This is the main goal of my project. Recently, a new model has been proposed to generate strong indicators for predicting price reactions to new information. This model [59] generated a powerful return-predictive signal from their supervised sentiment text model with news data. Our research methodology for generating sentiment scores referred to the methods suggested in [59]. However, our study will show an innovative improvement compared to [59]. [59] employed news articles only to generate return-predictive signals, while our study will use 10-K reports, especially focusing on the risk factor section, to generate volatility-predictive signals as well as return-predictive signals. Furthermore, the predictive sentiment signals will be generated in three different stakeholder levels such as a sector, a portfolio, and an individual firm. As far as we are aware, no research applies supervised sentiment learning with 10-K reports for predicting volatility, nor is there any that offers a comparative study of pre-established and acquired sentiments by using 10-K reports for returns or volatility predictions.

- Build an automated pipeline on the Airflow framework.

10-K fillings should be released annually, but the publication dates of it are different to each firm. In the case of 100 firms listed in the QQQ for our study, for instance, a firm's 10-K is released almost every single day within that year. Due to that, in order to offer prompt sentiment information, our system should update the latest 10-K filings daily or at least monthly.

- Evaluate sentiment scores in the context of prediction for returns and volatility.

After we achieve our main goal, we will evaluate our generated sentiment scores quantitatively and qualitatively. For quantitative analysis, we will use Pearson correlation

(check Appendix E for the formula) to find a correlation between the metrics we generated. For qualitative analysis, we will employ the top 15 most influential words we extracted from the process of sentiment score generation. The most impactful words will be used to generalise a topic which may affect a sentiment.

1.3 Contribution

we have completed the main goal, including all sub-goals, we set in the previous project objective section.

Our contributions are shown as follows:

- Developed an automated system for generating sentiment analysis metrics, comprising two main models: the 10-K Filing Extraction Model and the Sentiment Score Prediction Model.
  - Implemented the 10-K Filing Extraction Model to automatically retrieve 10-K filings from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, including the extraction of the Item 1A risk factor disclosure section. This data feeds into the subsequent Sentiment Score Prediction Model. Referenced in [99].
  - Implemented the Sentiment Score Prediction Model to evaluate sentiment across three levels of stakeholders: sector (e.g., Technology), portfolio (e.g., top 10 firms), and individual company (e.g., Nvidia). Referenced in [59]. Unlike the [59] model, our model demonstrated results across three levels of stakeholders to maximise the model's applicability.
    * Formulated 12 sentiment metrics across these levels:
      · **Sector Level:**
      · Developed sector sentiment metrics from 10-K filings, labelling based on QQQ's return and volatility.
      · Generated sector sentiment metrics from the Item 1A risk factor section, with labels for QQQ's return and volatility.
      · **Portfolio Level:**
      · Formulated portfolio sentiment metrics using 10-K filings, labelling based on portfolio returns and volatility.
      · Produced portfolio sentiment metrics from the Item 1A risk factor section, with labels for portfolio returns and volatility.
      · **Company Level:**
      · Established company-specific sentiment metrics through 10-K filings, labelling based on company returns and volatility.

· Derived company-specific sentiment metrics from the Item 1A risk factor section, with labels for company returns and volatility.

- Generated sentiment scores for 10-K/Item 1A by using [59]'s methodology. This approach is novel. No prior study has applied this methodology to 10-K filings/Item 1A risk factors.
- Utilised return/volatility to train sentiment scores. No other study employs such an approach. Even the [59] model did not use volatility to train sentiment signals. We produced volatility-predicative sentiment signals as well as return-predicative sentiment signals unlike [59].
- Assessed the performance of the Sentiment Score Prediction Model.
- Conducted both quantitative(i.e. correlation analysis) and qualitative analysis(i.e. most influential words to return/volatility) to critically evaluate the derived sentiment metrics. Through these evaluations, an investor can improve their understanding of the fundamentals of a market or a firm deeply.
- Connected to Airflow server for automation update. Our system can provide the latest sentiment scores for three stakeholder levels as this system automatically updates the latest 10-K fillings, facilitating seamless data preparation for the model. Investors can receive prompt sentiment information for a sector, a portfolio, and a firm. You can find detail explanation at Appendix H.

1.4 Outline of the Project

The thesis structure is outlined below:

- **Chapter 1 - Introduction** provided motivation for this study, including study objectives and contributions
- **Chapter 2 - Related Works** explained related works about various textual sentiment analysis approaches as well as its relevant background information.
- **Chapter 3 - 10-K filings Extraction Model** explained the 10-K filing acquisition process as well as the detailed extraction process of Item 1A risk factor section.
- **Chapter 4 - Methodology** detailed a comprehensive explanation of the sentiment score prediction model.
- **Chapter 5 - Experiment Results** critically evaluated all generated sentiment metrics through quantitative(i.e correlation analysis) and qualitative analysis(i.e most influential word set).
- **Chapter 6 - Conclusion** summarised the results of our suggested model and suggested to investors about model usage. Limitations and future works were mentioned.

Chapter 2

Related Works

2.1 Textual Sentiment Analysis in Finance

In the field of finance, sentiment analysis has grown in significance due to its wide range of applications. Through sentiment analysis, we can analyse, interpret, and derive insights from large volumes of financial data. One popular use case of sentiment analysis is stock market movement prediction [93, 100, 79, 43, 116, 102, 90]. In the context of sentiment analysis in stock market prediction, two types of sentiments have been researched: investor sentiment and textual sentiment. Investor sentiment refers to an investor's subjective sentiment on firms or markets. The second type of sentiment is textual sentiment or text-based sentiment. It indicates the level of positive or negative sentiment in texts. In some studies [59, 55], for instance, the term 'tone' (i.e. positive or negative) in a corporate disclosure means sentiment. Investor sentiment and textual sentiment are fundamentally different. Investor sentiment encompasses the subjective assessments and behavioural traits of investors. In contrast, textual sentiment may incorporate investment sentiment but also covers a more objective representation of the state within companies, institutions, and markets [60]. In the context of textual sentiment analysis, various sources have been used, such as news articles, social media data, or corporate disclosures. News articles and social media data are used to analyse the short-term effects of the sentiment on market variables like return, volatility, and stock volume [100, 116]. Corporate disclosure usually focuses on finding the relationship between sentiment and firm performance [46, 19]. In this paper, we focused on text-based sentiment using a corporate disclosure, analysing its impact not only on firm performance but also on portfolios and the market.

Furthermore, many previous studies to extract textual sentiments depended on pre-defined sentiment dictionaries instead of using statistical text analysis. Pre-defined sentiment dictionary is created through the dictionary-based approach. This approach utilises a mapping algorithm through which a computer programme reads text and categorises words, phrases, or sentences into predefined categories [71]. There are major pre-defined sentiment dictionaries for textual sentiment analysis in finance. The first one is the General Inquirer(GI) or Harvard IV-4 dictionary(HIV_4). The second one is the DICTION dictionary. The two dictionaries have been widely used for financial

analysis [110, 30, 31, 34, 48, 24]. However, both the GI/Harvard and DICTION, being general English linguistic dictionaries, offer inaccuracies in a financial context. To overcome this issue, Loughran and McDonald recreated the LM dictionary specific to the finance domain. Since the LM dictionary was developed for 10-K sentiment assessment, we used this dictionary as our benchmark to evaluate our estimated sentiments. The LM was used by many researchers [26, 52, 33, 20, 74].

2.2 Textual Sentiment Analysis with AI in Finance

With the increase in computing power and the development of cutting-edge AI methodologies, AI technology has been applied to textual sentiment analysis in finance. In classical machine learning, some previous papers utilised off-the-shelf machine learning techniques like Support Vector Machine(SVM), Naïve Bayes, Decision Trees, or artificial neural networks(ANN) to control the curse of high dimensionality in textual sentiment in finance context [51, 71, 72, 101, 115]. For instance, [115] extracted sentiments from social media textual data(e.g. StockTwits) through a variety of machine-learning binary classifiers. They found that the SVM classifier achieved higher accuracy than Decision Tress and Naïve Bayes. However, classical machine-learning techniques could not capture a sentence's complex features and contextual information. These tasks require deep-learning techniques, which facilitate understanding sequential information, complex feature extraction, and location identification [103].

Deep learning, a branch of machine learning, employs multi-layered neural networks, called deep neural networks, for extracting complex features [53]. In textual sentiment analysis, deep learning can be useful for generating learning patterns and learning contextual information in a sentence [124]. Many studies showed the effectiveness of deep learning models, such as current neural networks(RNN) [109, 108], convolutional neural networks(CNN) [62, 125, 56], and attention mechanisms [103, 121] for sentiment analysis in finance. Recently, transformer architectures such as BERT or RoBERTA have shown superior performance in sentiment analysis in the finance domain [81]. However, transformers' superior performance comes at a cost. They demand extensive data and computational power for training and testing. Moreover, they require considerable prediction times, rendering them less viable for real-time applications or environments with constrained processing capabilities [92]. Also, the transformer is perceived as a 'black box' due to its uninterpretable complex internal mechanisms [25].

Recently, [59] suggested an interesting and innovative sentiment model with a good performance. The suggested model is a lexicon learning model to find the correlation between the sentiments from firm-specific news and returns. This model has five virtues. First, this model is a transparent and simple supervised lexicon learning model. It needs only basic econometric methods, such as correlation analysis and maximum likelihood estimation. Hence, this approach is entirely 'white box'. Secondly, this model requires minimal computing resources. It only takes a matter of minutes to handle millions of documents on a laptop computer. Thirdly, this model has a broad range of scalability. Unlike existing lexicon-based models that depend on a pre-existing sentiment dictionary, this model is able to use various types of textual data in the finance domain without

a pre-defined dictionary. Fourthly, this supervised model does not require manual labelling labour to train the sentiment model. The model is free from the expensive expense of a significant amount of manual labelling labour. With minimal, reliable, and clever assumptions, the labelling mechanism works in an automated process. Finally, by training on returns from sampled companies to analyse sentiments, this method is likely more effective at predicting stock returns compared to others.

In this paper, we gained access to the model's other four benefits by leveraging its scalability. In other words, unlike the model using news articles to generate return-predicative sentiment signals, we used informative 10-K fillings(plus, Item 1A risk factor section) to produce volatility-predicative sentiment signals as well as return one instead. No prior study applied the model's methodology to the 10-K fillings and risk factor section. Also, even in this methodology, they did not use volatility to train sentiment signals. Moreover, we broadened the range of model application levels from a portfolio level to a sector level and a firm level to maximise the model's applicability.

Chapter 3

10-K filings Extraction Model

In this research, we utilised textual data for financial analysis. Among the myriad kinds of textual data available, we selected Form 10-K filings, which contain some of the richest information about firms. This paper collected the entire 10-K filings for predicting sentiment scores while collecting the Item 1A risk factor section of the filing to extract more informative features. Form 10-K filings are the official documents that all publicly traded firms in the United States are required to submit to the SEC. The SEC enforces very strict rules regarding the content and structure of the information required in Form 10-K filings. These filings contain no pictures or charts. A well-structured 10-K is divided into five separate parts. The first three parts offer a concise summary of the firm's primary business activities, including its services and products; enumerate every risk encountered by the firm; and provide detailed financial information about the firm over the last five years. The fourth part delivers a senior management analysis of its financial results. The final part includes the actual financial figures; the firm's audited financial statements, which consist of the income statement, balance sheets, and statement of cash flows. A detailed description of the 10-K structure can be found in Appendix A. Hence, the Form 10-K exhibits uniform structures. Thanks to these organised structures, we can algorithmically extract information on the filings through Figure 3.1. [49, 99]

3.1 10-K filings Collection

Form 10-K filings can officially be found on the SEC website, where they are available to the public. The SEC provides 10-K filings in HTML or TXT formats through its database system, known as the Electronic Data Gathering, Analysis, and Retrieval(EDGAR). This system offers various official filings, including 10-K, 10-Q, 8-K, and 6-K, among others. For our research, we were able to collect most of the 10-K filings from firms in the technology sector on the EDGAR. EDGAR offers files in both TXT and HTML formats, but since most of the 10-K reports were easily accessible in HTML format, our algorithms focused exclusively on extracting HTML files. For instance, EDGAR has 8140 filings in HTML format of all firms in the S&P 500, whereas, there are only 48 filings in TXT format [99]. We considered the portfolio of

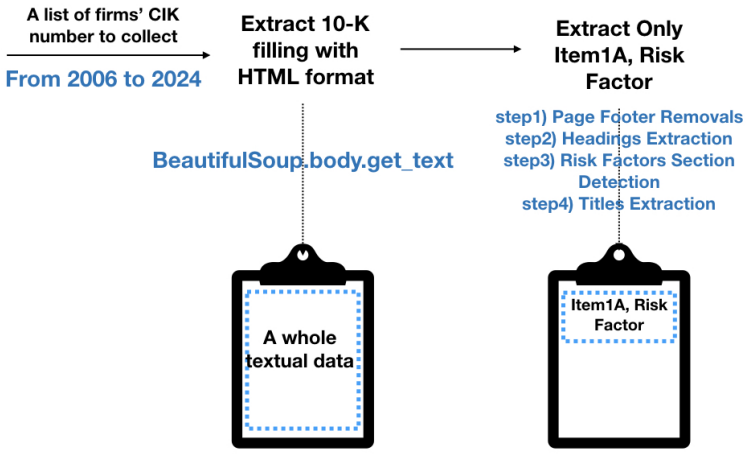

Figure 3.1: 10-K filing Extraction Model

Invesco QQQ Trust Series 1, an exchange-traded fund(QQQ or QQQ ETF), to compile a list of tech firms representing the technology sector in the United States. The QQQ is designed to passively track the Nasdaq 100 Index, which conventionally represents the technology sector. More details of the QQQ portfolio can be found in Portfolio Appendix C

Form 10-K filings 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 filing. Initially, we collected the CIKs list of firms listed in QQQ to facilitate data retrieval through the RSS Feeds provided by EDGAR. It is important to consider the crawler's limitation, which is a maximum request rate of ten requests per second, to ensure equitable access. In this report, we specifically focused on Item 1A, the risk factor, as well as the entirety of the 10-K filing for our purpose. As the mandatory disclosure of Item 1A as a separate section of the filing has been required since 2005, our crawler collected 10-K forms spanning nearly 17 years, from January 2006 to December 2023, for every firm listed in QQQ. In total, we collected 1383 filings, primarily stored in HTML. You can find an example case in Appendix B

3.2 Risk Factor Extraction Process

In this study, our focus was on Item 1A, the Risk Factor section, to attain richer information. Initially, we extracted all contents from the Item 1A Risk Factor section. We separately extracted each risk heading along with its corresponding content and then aggregated these elements together. To extract only the risk factor section, we used a well-organised structure of 10-K. You can find an example case in Appendix B

The Form 10-Ks feature well-structured formats in HTML. We observed the following

regularities in the structure of the 10-Ks:

- The heading of a risk is followed by the explanations of the risk (Check the example in Appendix B.
- There is a summary of the Risk Factor section at the beginning of the section.
- Non-informative elements are located in the footers, including page numbers, copyright information, or disclaimers
- Reiterated expressions appear in a certain section; specifically, "Item 1A Risk Factor (Continued)."
- Diverse layouts exist within a report, with variations in fonts, headings, and overall formatting.

With the aid of the regularity in the filing's structure in HTML, our algorithms can accurately locate and extract the risk factor section. The process of extracting risk factors from HTML involves four steps:

- Page Footers Removal
- Headings Extraction
- Risk Factor Section Detection
- Titles Extraction

3.2.1 Page Footers Removal

There is repetitive information in page footers of HTML files. Initially, we removed "Item IA. Risk Factors(Continued)" by using RegExr 4 at Table 3.1. We observed that page footers are typically found near horizontal lines marked by '<hr>' or ' ' tags with 'page-break-after: always' RegExr 2 at Table 3.1. They were then replaced with a 'split of pages' marker by using the BeautifulSoup library. Our algorithm subsequently identified and removed non-informative page footers- both textual and numerical- located near these markers by scanning both forward and backward.

3.2.2 Headings Extraction

Headings were extracted by detecting tags typically associated with headings, identified by their styles or attributes. The algorithm uses these attributes at Table 3.1 to identify headings and encapsulate their contents within ' <heading>' and ' </heading>' markers. These identified headings are then readily identifiable for further analysis steps. Additionally, we removed the ' <heading>' and ' </heading>' tags when the heading contained five words or fewer.

3.2.3 Risk Factor Section Detection

The risk factors section is extracted by identifying the heading "Item 1A - Risk Factor" RegExr 6,7 at Table 3.1, and the subsequent heading RegExr 8-16 at Table 3.1. The

contents of the section are located between the heading and the subsequent headings. Typically, if the pattern “Item 1A -Risk Factor” appears in the heading, it may contain extra spaces, line breaks or variations. Conversely, if it is not in the heading, the pattern is consistently enclosed in special quotation marks, either “ ldquo”, “ rdquo”, or “ quot”. The algorithm, improved by [49], detects the positions of the several types of headings by iterating through the regex pattern at Table 3.2. Additionally, after extracting the contents of the section, the algorithm checks if the extracted content typically exceeds 1000 characters in length.

3.2.4 Titles Extraction

The extracted headings include both titles (such as “Risks Related to Legal, Regulatory and Compliance Matters; FORWARD-LOOKING STATEMENT”) and headings (like “We may be unable to adequately protect our proprietary intellectual property rights, which may limit our ability to compete effectively.”). We noted that titles have all words starting with capital letters, after removing stopwords. Thus, for headings fitting this pattern, we replace their markers with ‘<title >’ and ‘</title >’.

ID	Description	Regular Expression
1	Removal of “Continued” pattern in page footer	ITEM 1A(.0,10)RISK TORS.0, 10(continued)
2	Search tags with specific style	page-break-after*:*always
3	Extraction of Item 1A.	[^”“]ITEM s*1A[^””]
4	Extraction of Item 1A. (alternative)	RISK[^a-zA-Z0-9”“]*FACTOR s*S[^””]
5	Extraction of Item 1B.	ITEM s*1B
6	Extraction of Item 1B. (alternative)	Unresolved*[^a-zA-Z0-9]*Staff*[^a-zA-Z09]*Comments
7	Extraction of Item 2.	ITEM s*2
8	Extraction of Item 3.	ITEM s*3
9	Extraction of Item 2. (alternative)	Legal*[^a-zA-Z0-9]*Proceedings
10	Extraction of special case ’Management’s Report’	Management*[^a-zA-Z0-9]*S[^a-zA-Z09]*Report
11	Extraction of special case ’Managing Global Risk’	Managing*[^a-zA-Z0-9]*Global[^a-zA-Z09]*Risk
12	Extraction of special case ’Statement of Income Analysis’	Statement*[^a-zA-Z0-9]*of[^a-zA-Z09]*Income[^a-zA-Z0-9]*Analysis
13	Extraction of special case ’non GAAP Financial Measure’	non*[^a-zA-Z0-9]*GAAP[^a-zA-Z09]*Financial[^a-zA-Z0-9]*Measure

Table 3.1: Regular expression for extraction in HTML

3.3 Preliminary Analysis

The data extraction algorithm successfully collected the 10-K filings from 94 firms out of 100 firms listed in the QQQ ETF directly from the SEC EDGAR database. The six firms that are not included are foreign-based and instead issued 20-F filings. In total, the algorithm gathered 1397 filings, demonstrating its effectiveness for 10-K filings. However, some documents do not follow the standard regularity, including having an unstructured format or placing the risk factor section in other sections. The number

Sample	Description
font-weight: bold	Bold heading with style attribute
font-weight: 700	Bold heading with style attribute
	Bold tag
¡strong¿¡/strong¿	Strong tag
text-decoration: underline	Underlined heading with style attribute
font-style: italic	Italic heading with style attribute
	Italic tag
	Emphasized tag

Table 3.2: Heading Pattern

of non-standardised documents is 63 ouf of the 1397 documents. Thus, we collected 1334 filings in total, excluding 63 fillings. In these cases, our algorithm was not able to accurately identify and collect the relevant information. Despite this fact, its design allows for adaptability to collect various types of SEC filings, making it a scalable tool for financial research that requires textual report analysis.

Chapter 4

Methodology

In this section, we introduced the supervised learning model for 10-K sentiment analysis and explained how the model functioned. Section 4.1, which was referred to as [59], demonstrated how the model generated a sentiment score of a 10-K filing by using the return at the time of the filing's publication as a label. In Section 4.2, we described in detail the model that used volatility as a label to estimate sentiment and its adaptation process. In Section 4.3, to represent the macroscopic sentiment trend of the technology sector, we introduced the Kalman filter and its process for removing noise in the context of 10-K sentiment analysis.

4.1 A Sentiment Score Prediction Model

4.1.1 Notation

To establish notation for the probabilistic sentiment model, we considered n as a set of 10-K filings and V as a vocabulary consisting of m words. The $i = 1,\ldots,n$ represented the index of 10-K filings. It represented both the filing's publication date and the company to which it related. The word counts were recorded in a vector $d_i \in R_+^m$, where $d_{i,j}$ represented the number of times word j occurred in 10-K filing i. We defined $D \in R_+^{n\times m}$ as a document-term matrix, with $D = [d_1, ..., d_n]$, representing word counts in each document. d_i was the i-th row of D, and the indices of columns were listed in the set S, which was a subset of vocabulary, i.e., $S \subseteq V$. $D_{[S],i}$ was the submatrix of the i-th filing. $d_{[S],i}$ was the word count vector in subset S for the i-th filing.

We labelled filing i with the associated time series variable y_i, either return or volatility, on the publication date of the filing. In this project, we assumed that each filing had a sentiment score, denoted by $p_i \in [0,1]$. In the case of return fitting, a high score suggested that the filing had a predominantly positive tone in the report. However, it implies neither positive nor negative in the context of volatility fitting, as volatility signifies market uncertainty. Therefore, in the volatility setting, a high score indicates high market uncertainty and a low score suggests low market uncertainty.

4.1.2 Model Assumptions

4.1.2.1 Assumption 1

We assumed that vocabulary V consisted of a set S of sentiment-charged words and a set N of neutral words, i.e., $V = S \cup N$. These sets were mutually exclusive, i.e. $S \cap N = \emptyset$. Furthermore, we posited the set of sentiment-charged words affected the tone of a filing, whereas the set of neutral words did not influence its tone, i.e, $d_{[S],i} \not\perp\!\!\!\perp p_i$ and $d_{[N],i} \perp\!\!\!\perp p_i$. The sentiment word count was independent of the neural word count, i.e., $d_{[S],i} \perp\!\!\!\perp d_{[N],i}$, implying that the model did not include the neutral words.

4.1.2.2 Assumption 2

We assumed that the sentiment-charged word counts $d_{[S],i}$ were produced by a mixture multinomial distribution:

$$d_{[S],i} \sim \text{Multinomial}\left(s_i, p_i O^+ + (1 - p_i) O^-\right) \tag{4.1}$$

,where s_i was the total count of sentiment-charged words in the i-th filing, i.e., $s_i = \sum_{w \in S} d_{w,i}$. This determined the scale of the multinomial. We then modelled $O \in R_+^{|S| \times 2}$ matrix, representing the probabilities of individual word counts using a mixture model that incorporated positive and negative topics. The model included O_+, a vector of $|S|$ non-negative elements with a unit ℓ^1-norm, representing a probability distribution across words, such that $\sum^{|S|} o_{+,w} = 1$, where $o_{+,w} \in O_+$ and $w \in |S|$. O_+ symbolised a 'positive sentiment topic' and represented the expected distribution of word frequencies in a filing with the highest possible positive sentiment, where the probability p_i equals 1. Similarly, O_- symbolised a 'negative sentiment topic' that represented the distribution of word frequencies in a filing with the most pronounced negative sentiment, where the probability p_i equals 0. The sentiment score p_i, where $0 < p_i < 1$, was a mixture coefficient of two sentiment topics.

4.1.2.3 Assumption 3

Finally, we assumed that the sentiment score fully encapsulated the information within a filing that affected the dependent variables, i.e., $y_i | p_i \perp\!\!\!\perp d_i$. This implied that sentiment score could primarily be used as a feature for return or volatility prediction models. $\check{p}$

4.1.3 Model

The model, incorporating these three assumptions, consisted of three steps for predicting the sentiment score of a 10-K filing. First, we extracted the set of sentiment-charged words, denoted as $\hat{S}_n$. Second, we estimated the probabilistic distribution matrix of positive-negative topic parameters over words, which is $\hat{O} = [\hat{O_+}, \hat{O_-}]$. Third, we predicted the sentiment scores $\hat{p_i}$ of a 10-K filing using penalised maximum likelihood estimation. Each step was described in detail in Section 4.1.3 to Section 4.1.4. The model overview can be found in Figure 4.1.

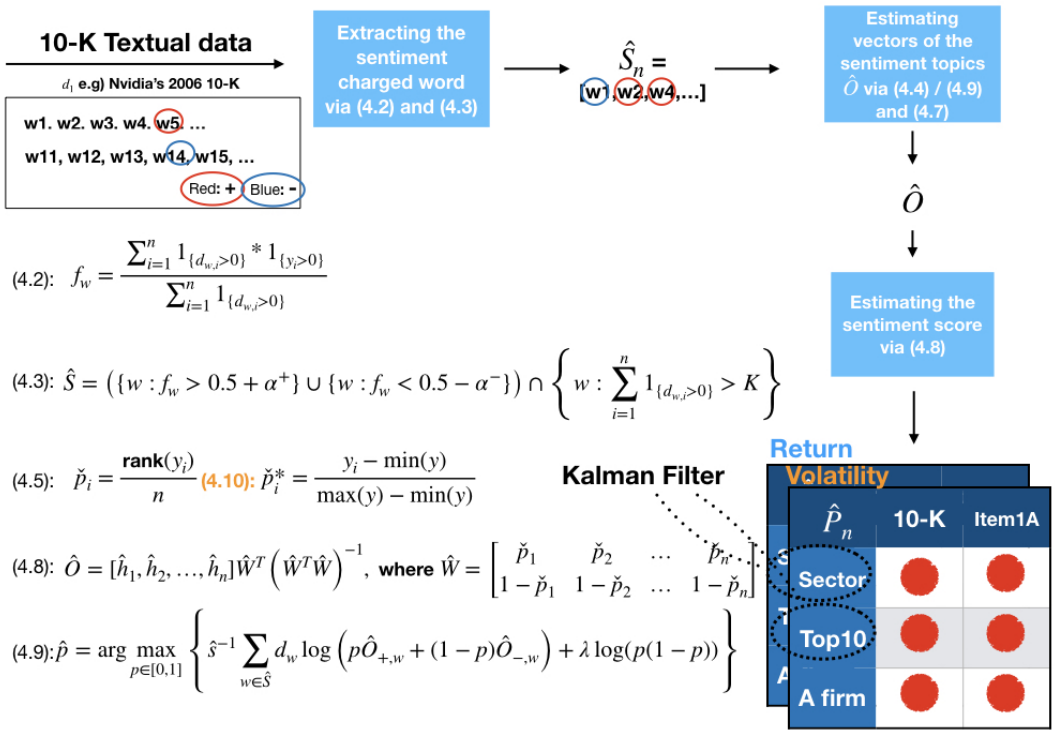

Figure 4.1: A Sentiment Prediction Model

4.1.3.1 Extracting sentiment-charged words

In a 10-K filing, sentiment-neutral words were likely to predominate in terms of the number of words and total counts. This dominance tended to introduce noise and could make the extraction of sentiment-charged words from the entire vocabulary computationally burdensome, especially if sentiment-neutral words were not selectively excluded. Hence, in our model, we filtered out the set of sentiment-neutral words and focused solely on the subset of sentiment-charged words, estimating topic parameters for this subset alone. To achieve this, we utilised realised stock returns as a label (Check Return y_i, or R_t calculation) (Volatility was also used as a label, with the fitting process described in Section 4.2 in detail). The indicator function, represented by $1_{\{a\}}$, was defined as 1 when condition a was met, and 0 otherwise. Consequently, $1_{\{d_{i,w}>0\}}$ represented the presence of word w in the i-th filing, and $1_{\{y_i>0\}}$ denoted that the return variable associated with the i-th filing was positive. For each word $w \in V$, we could compute the frequency of word w in 10-K filings with positive returns relative to its overall frequency across all filings using the following equation:

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

This measure signified the word's sentiment tone to return values. For any given word w, if filings that contained it generally coincide with positive rather than negative returns,

w was tagged with a positive sentiment. Conversely, if occurrences of w tended more often to align with negative returns, it was considered to have a negative return.

Subsequently, we assessed f_w against appropriate thresholds. If f_w was approximately 0.5, this suggested that the word was sentiment-neutral and belonged to the set N. To differentiate sentiment-charged words, we defined α_+ and α_- within the interval $(0, 0.5]$ to filter out sentiment-neutral words. A word was classified as a positive sentiment word if $f_w > 0.5 + \alpha_+$ and as a negative sentiment word if $f_w < 0.5 + \alpha_-$. We also defined a third threshold, $k \in N$, which pertains to the count of word w across all filings. The threshold k acted as a minimum frequency requirement to mitigate the impact of rare words, which could be sources of noise. For instance, a term appearing only once in all filings would result in an f_w of either 0 or 1, thus being noisy and unreliable itself. By applying the threshold k, we filtered out such noisy instances, requiring that a word appeared more than k times to be considered: $\sum_{i=1}^{n} 1_{\{d_{w,i}>0\}} > k$. With these conditions, our extracted set $\hat{S}$ was defined by

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

***Return y_i, or R_t calculation**

When using a return as our label, we used the publication time t over the days $t-1$ to $t+1$, as suggested in [59]. Using only the return at the publication time t can produce a noisy signal. A return may slowly respond to the content of 10-K filings, so the market may need the time to reflect the released 10-K filing of a firm to its stock price. Additionally, some contents of 10-K filings can already be reflected in its return as the 10-K filings, annually released, can contain some repetitive contents. To mitigate the noisy signal, we used open-close log returns R_t:

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

where P refers to the equity price at a given time. The market open time at day t was denoted t_O, and the market close time at day t was denoted t_C. We used open-close return to make a symmetric alignment with the 10-K filing's published date.

4.1.3.2 Estimating probabilistic distribution of topic parameters

In this process, our goal was to estimate the topic parameters for a report. $\hat{O}_i = [\hat{O_i}+, \hat{O}_-]$ referred to a 1x2 vector, with each element corresponding to the estimated probabilistic distribution of positive topic or negative topic for a filing, respectively. To obtain $\hat{O} = [\hat{O}_+, \hat{O}_-]$, we associated the sentiment expressed in a 10-K filing with stock returns or volatility (Fitting on volatility will be explained on Section 4.2). This approach comes from the assumption that stock returns or volatility at the publication date of a 10-K filing represented the sentiment for it. Note that we did not have direct

sentiment scores for the filings, thus we estimated the sentiment proxy by using the standardised ranks of stock returns as a label in the equation:

$$\check{p}_i = \frac{\text{rank}(y_i)}{n}, \tag{4.5}$$

where the expression defined $\check{p}_i$, the estimated sentiment proxy for the i-th filing, as the rank of y_i, i.e, a stock return or volatility (Fitting on volatility will be explained on Section 4.2), divided by n, the total number of filings. The rank of y_i was sorted in ascending order.

With these estimated sentiment proxies, we had a matrix $\hat{W}$ containing two rows for each filing: one for the estimated positive sentiment proxy $\check{p}_i$ and one for the estimated negative sentiment proxy $1-\check{p}_i$ in the equation:

$$\hat{W} = \begin{bmatrix} \check{p}_1 & \check{p}_2 & \cdots & \check{p}_n \\ 1-\check{p}_1 & 1-\check{p}_2 & \ldots & 1-\check{p}_n \end{bmatrix}, \tag{4.6}$$

Subsequently, we adjusted the counts of sentiment-charged words by dividing each count by the total sentiment-charged word count for the filing like in the equation:

$$\hat{h}_i = \frac{d_{[\hat{S}],i}}{\hat{s}_i}, \text{ where } \hat{s}_i = \sum_{w \in \hat{S}} d_{w,i}, \tag{4.7}$$

where $\hat{h}_i$ was the counts of sentiment-charged words. This was defined as the ratio of $d_{[\hat{S}],i}$, each word count within the subset $[\hat{S}]$ of the i-th filing, to $\hat{s}_i$, which was the total sentiment-charged word count for the filing. These relative term counts were collected in a matrix $\hat{H} = [\hat{h_1}, ..., \hat{h_n}]$.

In this process, we aimed to estimate a matrix $\hat{O}$, which contained parameters that referred to the probability distribution of positive and negative sentiments. Then, we could estimate $\hat{O}$ with a regression, using matrix $\hat{H}$ as the predicted outcome and matrix $\hat{W}$ as the predictor like in the equation:

$$\hat{O} = \hat{H}\hat{W}^T(\hat{W}\hat{W}^T)^{-1}. \tag{4.8}$$

In the final step, to ensure the estimates correspond to a probability distribution, we corrected any negative entries by resetting them to zero and then re-normalised each column so that their totals were equal to one. This process produced a revised matrix, but to simplify notation, we reused $\hat{O}$ for the resulting matrix, and we labelled its first and second columns as $\hat{O_+}$ and $\hat{O_-}$, respectively.

4.1.4 Scoring New Filing

Now that we constructed estimators $\hat{S}$ and $\hat{O}$. Also, from the mixed multinomial distribution in Assumption 4.1.2.2, we could estimate the filing's count vector, which is

$d_{[S]}$. Given all estimates $\hat{S}$, $\hat{O}$, and $d_{[S]}$, we could estimate the best probable sentiment score p by Maximum Likelihood Estimation (MLE):

$$\hat{p} = \underset{p\in[0,1]}{\arg\max} \left\{ \hat{s}^{-1} \sum_{w\in\hat{S}} d_{i,w} \log(p\hat{O}_{+,w} + (1-p)\hat{O}_{-,w}) + \lambda \log(p(1-p)) \right\}, \quad (4.9)$$

where $\hat{s}$ represented the total count of words from the set $\hat{S}$ in the new filing, while $d_{i,w}$, $\hat{O_{+,w}}$, and $\hat{O_{-,w}}$ referred to the w-th elements of their respective vectors, and λ was a positive constant used to adjust the model. In the MLE, $\lambda \log(p(1-p))$ was a penalty term to avoid overfitting when there were reports with few sentiment-charged words. For instance, if the filing had a limited number of positive sentiment-charged words without negative words, the model believed the filing had a positive tone even if the filing just only contained a few positive words. The term served as a regularising factor that pushed the estimated sentiments toward 0.5, indicative of a neutral sentiment. In other words, the penalty terms nudged the model to be more conservative when scoring new filings.

4.2 Volatility Label

Section 4.1 introduced the sentiment prediction model with the firm return as y_i, as in [59]. Notably, [59] suggested the model is universally adaptable. However, its core assumptions and methodologies might not suit every forecasting objective such as using volatility as a label. While the current model with a return label fits into a binary or discrete framework, volatility does not naturally fit into a binary or discrete framework. Due to that, we should do adaptation to use volatility as a label for the model.

In the current narrative of returns prediction, we could employ returns as a label corresponding to binary sentiment topics, i.e., positive and negative, because one either made a loss or a profit. However, this classification was not clear in the context of volatility. The adaptation for volatility was to set a threshold θ and we labelled all values above this θ point as high volatility and those below it as low volatility. We then replaced 0 by the θ in Equation 4.2 and the value 0.5 by quantile q in Equation 4.3. Note that there were no right threshold or quantile, but we should keep in mind that our choice of hyper-parameter would impact the outcome of the model.

Volatility is, in nature, an asymmetric variable; thus, getting a ranking of the volatility like in Equation 4.4 will lose substantial informational value to estimate the sentiment score. Subsequently, normalising the volatility can be an appropriate alternative as it preserves asymmetry:

$$\hat{p}_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}, \quad (4.10)$$

where $\hat{p}_i^*$ is greater than or equal to 0 and less than or equal to 1. However, the normalised volatility itself was not able to catch the outlier of the market movement

despite making it symmetric around zero, leading to poor predictive performance. Thus, we used volatility values over multiple days for the robust labels. In practice, we averaged the volatility over three days. We used the intra-daily range among other standard volatility proxies such as squared returns, or the realised volatility as it is a less noisy volatility proxy, leading to less distortion [6, 86]. The intra-daily log range is defined as:

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

The volatility proxy $\tilde{V}_t$ was then computed like this:

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

where the intra-daily log range was squared and divided by the adjustment factor, $\frac{1}{4\log(2)}$. The adjustment factor is used to correct potential bias that may occur in the data generation process when we assume that the process follows Brownian motion with drift Parkinson, 1980. Then, we attained the averaged volatility over three days for a more reliable label:

$$V_t = \frac{1}{3}\left(\tilde{V}_{t-1} + \tilde{V}_t + \tilde{V}_{t+1}\right), \tag{4.13}$$

which is the common approach to calculating multi-day aggregation of daily volatility proxies [22].

4.3 Kalman Filter

In this paper, we generated sentiment metrics of both the technology sector and a single firm(e.g. Nvidia) with 10-K filings over a certain time. But, the estimation of the industry's time-varying sentiment measures should be very noisy. To obtain robust sentiment features on time series, we adapted the Kalman filter to smooth the time-varying noisy signals. In the context of smoothing sentiments of 10-K filings across the industry, we referred to [13] to adapt the Kalman filter to our context. Following the adapted Kalman filter in our context of work,

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{4.14}$$

$$v_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{4.15}$$

Firstly, we assumed that there were the observed sentiment score
v_t and an unobserved sentiment score - the state variable μ_t in the Kalman filter framework. We extracted the unobserved sentiment score from the publication-date-based sentiment scores via Equation 4.14 and Equation 4.15. Equation 4.14 referred to the

prediction model and Equation 4.15 referred to the correction model. The prediction model updated the unobserved state, and the correction model represented the observed sentiment cores that were affected by the state variable. The Kalman filter worked recursively through prediction and correction. It predicted the unobserved state estimates from the previous weighted variable. You can find a more detailed explanation for this Kalman filter model in [27]. In our study, we employed the Kalman smoother, instead of the Kalman filter, as the former one was likely to be more proper to show a retrospective trend to analyse the industry-level sentiments. Both the Kalman filter and the Kalman smoother are very similar in terms of estimating the unobserved sentiment score. However, the Kalman filter estimates the unobserved sentiment score at time t given observations up to and including time t, whereas the Kalman smoother estimates the hidden sentiment score based on all observations for $t = 1, ..., T$. In this paper, we called the Kalman smoother as the Kalman filter for convenience.

To adapt the Kalman filter in our context, our sentiment scores should be converted to time series data, because there were multiple sentiment scores on the same publication data of 10-K filings. In order to convert it to time series data, we aggregated sentiment scores on the same date over all firms like $\bar{p}_t = \frac{1}{|A_t|} \sum_{i \in A_t} \hat{p}_i$, where A_t referred to the set of 10-K filings published on day t. Subsequently, v_t was substituted by $\bar{p}_t$ in Equation 4.9 and we obtained the filtered industry-level sentiment scores $\tilde{p}_t$ after applying the Kalman filter. In other words, this approach implied that all 10-K filings published on the same day were treated equally. This approach may enhance the model's ability to estimate sentiment scores more accurately. However, it comes with the trade-off of transforming the score into an aggregate metric of all 10-K publications within a given day.

Chapter 5

Experiment Results

In this paper, we aimed to generate various sentiment scores from our suggested sentiment prediction model. To achieve that, we generated sentiment scores at three stackholder levels: Sector level, Portfolio level, and Company level. At the sector level, we generated scores from the model where it aggregated all filings of companies from the technology sector. At the portfolio level, we generated scores for a specific set of companies. In the practice of our study, we aggregated filings of the top 10 firms listed in QQQ. Finally, at the company level, we aggregated all filings of a firm to generate sentiment scores. We generated Nvidia's sentiment metrics in our study. All models for each level commonly used either the entire 10-K or only the risk factor section, by labelling with return or volatility. That is each level generated four types of sentiment scores. Three levels were multiplied by four types of sentiment scores. In total, we generated 12 types of sentiment scores(check Contribution). Furthermore, for analysis of both a sector level and portfolio level, we applied the Kalman filter to mitigate noise.

In this chapter, we showed the statistics overview of experimental results to introduce what we found. Then, we evaluated sentiment scores we generated at three stack holder levels through correlation analysis and qualitative analysis with the most influential words.

5.1 Descriptive Statistics

This section shows descriptive statistics of sentiment scores we generated at the three different stakeholder levels. We show the number of observations for each model and the number of sentiment words used for each model. We also calculate the mean, Standard Deviation(SD), and maximum and minimum sentiment scores for each model.

Sector Level (i.e., Technology Sector)

	# of Total Tokens (# of Sentiment Tokens)	Mean	SD	Max/Min
$\tilde{p}^{RET}$ with 10-K (Figure 5.1)	15,863 (8,725)	0.48	0.28	0.92/0.08
$\tilde{p}^{VOL}$ with 10-K (Figure 5.1)	15,863 (9,518)	0.43	0.24	0.92/0.08
$\tilde{p}^{RET}$ with Risk factor (Figure F.1)	9,030 (5,147)	0.46	0.27	0.92/0.08
$\tilde{p}^{VOL}$ with Risk factor (Figure F.1)	9,030 (3,792)	0.42	0.25	0.92/0.08

Portfolio Level (i.e., Top10 Firms Portfolio)

	# of Total Tokens (# of Sentiment Tokens)	Mean	SD	Max/Min
$\tilde{p}^{RET}$ with 10-K (Figure F.2)	7,040 (6,434)	0.42	0.30	0.92/0.08
$\tilde{p}^{VOL}$ with 10-K (Figure F.2)	7,040 (6,985)	0.43	0.33	0.89/0.08
$\tilde{p}^{RET}$ with Risk factor (Figure F.3)	4,265 (4,052)	0.42	0.29	0.92/0.08
$\tilde{p}^{VOL}$ with Risk factor (Figure F.3)	4,265 (4,005)	0.44	0.31	0.88/0.08

Company Level (i.e., Nvidia)

	# of Total Tokens (# of Sentiment Tokens)	Mean	SD	Max/Min
$\hat{p}^{RET}$ with 10-K (Figure F.4)	3,861 (3,861)	0.49	0.32	0.89/0.13
$\hat{p}^{VOL}$ with 10-K (Figure F.4)	3,861 (3,861)	0.39	0.21	0.86/0.13
$\hat{p}^{RET}$ with Risk factor (Figure F.5)	2,201 (2,201)	0.49	0.32	0.90/0.11
$\hat{p}^{VOL}$ with Risk factor (Figure F.5)	2,201 (2,201)	0.41	0.18	0.80/0.17

5.2 Evaluation Methods

5.2.1 Correlation Analysis

We used the Pearson correlation to evaluate our estimated sentiment scores. The Pearson correlation coefficient could have two assumptions[105]. The first assumption is two variables are continuous. Secondly, two variables are assumed to have a linear relation. Based on these assumptions, we could analyse our sentiment scores with the Pearson correlation. We used four variables(i.e. P^{RET}, P^{VOL}, P^{LM}, and $Stock$) for correlation analysis. The correlation of variable pairs was calculated through the Pearson correlation formula in Appendix E, and their linear correlation could be checked through a significance test [82]. In a significance test, 0.05 is conventionally used as a significance level. However, a long list of scientists recently proposed a new p-value threshold(i.e. 0.005) to improve the reproducibility as the conventionally used threshold leads to a high rate of considerable false positives, even without taking into account any issues related to the experiment, procedure, or documentation [10]. In the practice of our study, we significantly increased our experiment's reliability by setting both 0.05 and 0.005.

5.2.2 Qualitative Analysis with Most Influential Words

In the process of predicting sentiment scores for all models we have introduced so far, we could extract the top 15 influential words from $\tilde{p}^{RET}$, and $\tilde{p}^{VOL}$for each model. The detailed extraction process can be found in Equation 4.3. With the extracted set, we could infer a topic for the analysis of three stakeholder levels.

When interpreting the set of the extracted words, the interpretations might be interpreted in various ways. The objective of the word extractions was to offer the most influential keywords used for financial analysis. We assumed the value of the words would be synergised with domain knowledge. Also, using generative AI could offer a good reference to our word set for in-depth and insightful analysis.

5.3 Evaluation Results

5.3.1 Sector Level (i.e. Technology Sector)

In a technology sector analysis, we computed the sector sentiment scores with all firms' filing, additionally considering the entire 10-K filing or only the Item 1A section, respectively. To improve readability, we introduced a table at the top right for all the score graph figures to identify easily which models you are referring to. Moreover, we employed a Kalman filter to represent a reliable sentiment trend for a sector analysis [27]. An industry-level analysis generated a considerable amount of signal noise. We could control these noisy signals by using the Kalman filter.

5.3.1.1 Sector Correlation Analysis

The Sector Sentiment Model with 10-K filing (Figure 5.1)

Figure 5.1 showed the predicted sentiment scores for the technology sector. These scores were estimated with almost all firms' 10-K filings listed in the QQQ(i.e. 94 firms out of 100), and we used the entire 10-K filings to calculate the scores. Note that the sentiment scores labelled with return and volatility in Figure 5.1 are filtered. Upon reviewing graphs in Figure G.1 and Figure G.2, it was evident that both unfiltered return sentiment and volatility sentiment contained significant noise. Therefore, filtered versions were chosen to more reliably depict the sector's trend. Other models, which will be introduced in the following parts, also used the filtered one.

We calculated two types of average loss for the same windows: one between the estimated sentiment score $\tilde{p}^{RET}$ and the normalized rank of *return*, and the other between the estimated sentiment score $\tilde{p}^{VOL}$ and the normalized rank of $volatility$. Furthermore, we calculated how well our sentiment analysis models can predict the sentiment of the financial markets based on return or volatility. For this model (Figure 5.1), the loss of the model $\tilde{p}^{RET}$ was 0.25, and the accuracy rate was 78%. It showed $\tilde{p}^{RET}$was strongly well predicted compared to the QQQ *return* given a window. It meant the $\tilde{p}^{RET}$ represented the sentiment of the technology sector. Additionally, the $\tilde{p}^{VOL}$ sentiment of this model showed stronger prediction accuracy. The $\tilde{p}^{VOL}$ model performed a 92% accuracy rate with 0.22 loss.

To evaluate our model critically, we selected the LM as our benchmark as it was created for 10-K sentiment assessment. In Figure 5.1, the $\tilde{p}^{LM}$ score showed a steady decrease until 2018, followed by a significant and gradual decline, with intermittent fluctuations due to noise. This trend occurred despite the technology sector's rapid growth from 2018 until just before the COVID-19 pandemic, indicating a generally negative direction in the model's movements. This was because the predominance of negative over positive vocabulary in the LM dictionary (2,355 negatives vs 354 positives out of 86,531 words) suggested a bias towards negative sentiment in the $\tilde{p}^{LM}$ score, as approximately 85% of the relevant vocabulary is negative. The rest, 83,822 words, were neutral and excluded from the sentiment analysis. This imbalance likely caused the $\tilde{p}^{LM}$ score to trend negatively. Additionally, the sector sentiment prediction model took 37 seconds to execute.

These tables (Table G.1, and Table 5.1) represented Pearson correlation coefficients between the sentiment estimates including QQQ's stock price, both filtered and unfiltered. Upon analyzing Table G.1, we found that, with the exception of the VOL and LM pair, there was no linear correlation between any pairs of unfiltered sentiment scores. However, analysis of filtered sentiment scores (Table 5.1) revealed more significant correlations compared to unfiltered scores. All other pairs, with the exception of the LM and $Stock$, exhibited weak correlations. Specifically, an r value of -.245 between $\tilde{p}^{RET}$ and $\tilde{p}^{VOL}$ indicated a weak negative correlation, suggesting that positive sentiment in RET generally corresponded to negative sentiment in VOL, and vice versa. Furthermore, an r value of +.296 between $\tilde{p}^{RET}$ and $Stock$ implies a weak positive correlation, indicating that positive sector sentiment in RET was associated with rising QQQ stock prices, and vice versa. In Figure 5.1, despite oscillated fluctuations due to noise from the sector, $\tilde{p}^{RET}$ showed a slow increase during the observed window. Notably, the sentiment trend in 2023 for $\tilde{p}^{RET}$ closely mirrored the QQQ stock performance in the same year. The case of $\tilde{p}^{VOL}$ and $Stock$ (r=+.121) showed also a positive correlation, but weaker. Moreover, both $r_{\tilde{p}^{RET},\tilde{p}^{LM}}$(=-.359) and $r_{\tilde{p}^{VOL},\tilde{p}^{LM}}$(=-.173) showed a weak negative correlation.

In both Table G.1 and Table 5.1, all pairs exhibited low p-values, with those in Table 5.1 being significantly lower. This suggests that the sample results—specifically, the correlation coefficients—provide sufficient evidence to reject the null hypothesis from the entire population [29]. In our case, the null hypothesis was that there was no correlation between the pair, whereas the alternative hypothesis was that there was a correlation between them. However, the lower p-value of all pairs does not measure the probability that the alternative hypothesis is true. The p-value is not an absolute index for arguing that a hypothesis is true. Instead, the lower p-value shows our experiments have statistical significance [37, 8, 85].

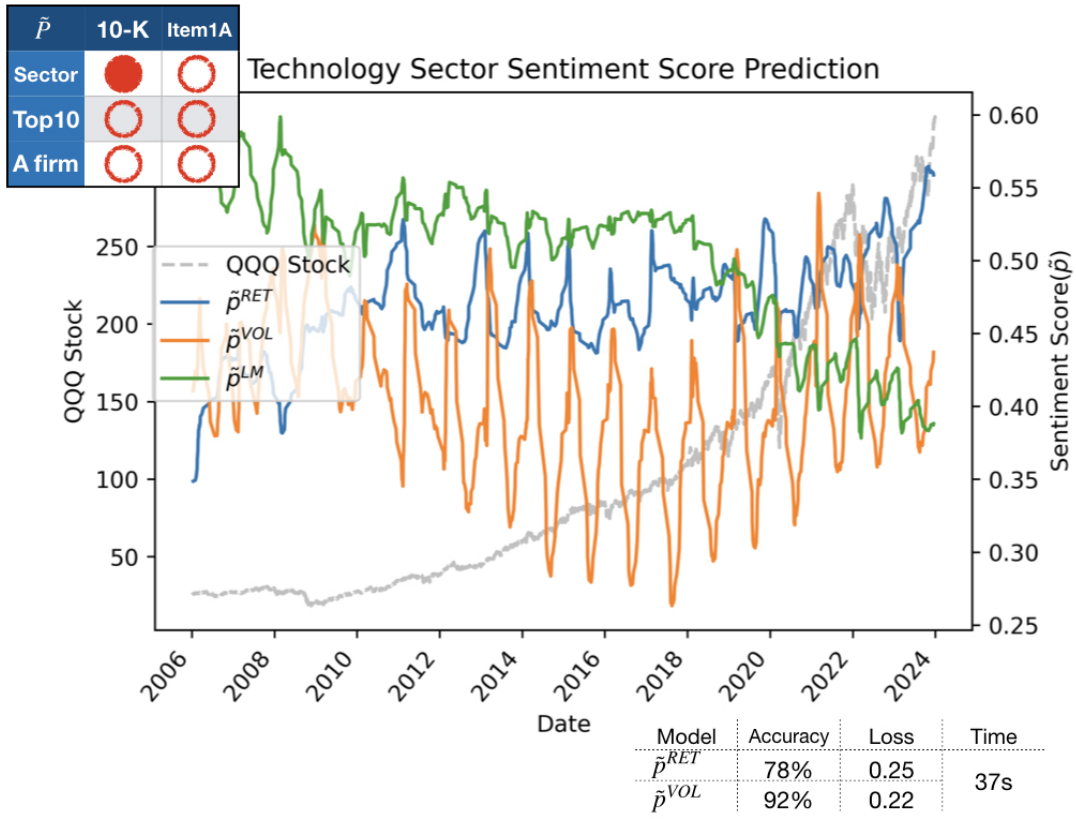

Figure 5.1: The Sector Sentiment Model with 10-K filing

$r_{\tilde{p}_i,\tilde{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	**-0.245	1		
LM	**-0.359	**-0.173	1	
Stock	**0.296	**0.121	**-0.910	1

Note :* $p\text{-value} < 0.05$,** $p\text{-value} < 0.005$

Table 5.1: Filtered, A Sector Sentiment Correlation with 10-K filing

The Sector Sentiment Model with Only-Risk-Factor (Figure F.1)

The sentiment scores in Figure F.1 were predicted based on the risk factor section. Note again that Figure F.1 only showed the filtered one (the unfiltered can be found at Figure G.3 and Figure G.4 in Appendix G).

The $\tilde{P}^{RET}$ model accuracy was 73% with the 0.25 loss. The risk factor $\tilde{P}^{RET}$ model accuracy is 5% lower than the entire 10-K $\tilde{P}^{RET}$ model, which is 78%. Moreover, the risk factor $\tilde{P}^{VOL}$ model showed a stronger model performance. It performed 90% accuracy with a 0.22 loss. Comparing the 10-K sector $\tilde{P}^{LM}$ and the risk factor $\tilde{P}^{LM}$, the latter is generally lower than the former. This pattern suggested that the $\tilde{p}^{LM}$ score, influenced by an LM dictionary dominated by negative words and a risk factor section pessimistically toned, inherently leaned towards a negative sentiment [15, 36].

When comparing the training data for both models(Figure 5.1, and Figure F.1), it's

noted that the aggregated word count from all QQQ firms' filings was 15,863 post-preprocessing, and the count from just the risk factor sections of these filings was 9,030 after undergoing the same preprocessing. This observation highlighted that the risk factor sections, despite typically constituting only 10% to 15% of the entire filing, contained a significant portion of the relevant words to the model's sentiments. It showed that a risk factor section could offer informational value for textual analysis in the finance domain. Also. the minor differences in accuracy between the models could suggest that the risk factor section provided a valuable textual context in predicting sentiment scores for both $\tilde{P}^{RET}$ and $\tilde{P}^{VOL}$. However, upon comparing both models, it was observed that all scores from the model analyzing risk factor sections were lower than those from the comprehensive model. It indicated that the tone of a risk factor section is pessimistic/negative as other studies also empirically found [15, 36]. The training time for this model, additionally, was also reduced by 40%(i.e. 23s)

Table F.1 showed the Pearson correlation coefficients of the filtered sentiment sector model with only the risk factor. Compared to the sector model with the entire 10-K filing, shown in Table 5.1, the sector risk factor model showed, in general, lower r values for all pairs, and some pairs showed altered sign such as the pair of $\tilde{P}^{RET}$ and $\tilde{P}^{LM}$, the pair of $\tilde{P}^{RET}$ and $Stock$, and the pair of $\tilde{P}^{VOL}$ and $\tilde{P}^{LM}$. Except for the pair of $\tilde{P}^{RET}$ and $\tilde{P}^{LM}$ which did not show statistical significance, the r=-0.376 from $\tilde{P}^{RET}$ and $Stock$ showed a weak negative correlation in the risk factor model. The more positive tone the sector has, the lower the price the sector has. This outcome was in contrast to that of the model (Table 5.1) and also our intuition where a good sentiment is related to a higher return. Other pairs did not show a meaningful correlation as they were close to 0.

5.3.1.2 Sector Most Influential Words

The Sector Sentiment Model with 10-K filing (Figure 5.1)
The following words were the most influential words in $\tilde{p}^{RET}$ sentiment score prediction, positively or negatively, respectively.

$$\tilde{S_+}^{RET} = \textit{underpay, secondly, dash, musk, incisive, memoranda, stance, maximizer, bolivia, torque, gilt, div, multifaceted, searcher, formulaic}$$
$$\tilde{S_-}^{RET} = \textit{transformer, scalar, tango, analgesic, invocation, carney, coconut, spectral, heath, rigorously, reimagine, infer, tera, markman, enclosure}$$

From the positive influential words in $\tilde{S_+}^{RET}$, we could categorise a few themes that could impact the technology sector's positive sentiment tone. The first theme can be 'Innovation and Development'. The words such as *musk, maximiser, searcher, multi-faceted*, and *incisive* could be interpreted as innovation-relevant vocabulary. *Maximiser* and *searcher* could refer to a figure to lead technology innovation or development. Interestingly, the word, *musk*, seemed to indicate "Elon Musk", who is one of the figures to represent innovation under the fourth industrial revolution era. Moreover, the words(i.e. *multifaceted, incisive*) could indicate a capability or capacity(or both)

required for innovation. In Figure 5.1, the technology sector has been gradually and significantly developed during the given period. This remarkable development could be associated with innovations. In this context, words such as *musk, torque,* and *bolivia* were additionally associated with innovation, specifically indicating an electric car or robotics. Elon *musk*, CEO of Tesla, leads the American firm known for its electric vehicles and robotics innovations. The *torque* word could be related to an electric vehicle or robot. *Bolivia* is home to the world's largest lithium deposits [17]. Lithium is a key component of electric car batteries. Furthermore, the words, including *gilt*, *div*(we interpreted as dividend), *memoranda*, and *stance*, could be interpreted as finance-relevant terms. Good financial health could be an important factor in economic growth [64]. Hence, these finance terms represent 'Financial Health'. The emphasis on improving the financial health of each tech firm might impact their stock positively, leading to a rise in the sector return.

We could infer a theme from the negative impactful words in $\tilde{S_-}^{RET}$. 'Technological Difficulty' could be a theme that makes the sector tone negatively. The words, such as *tango, transformer, scalar, infer*, and *tera* could represent a firm's technological difficulties. For instance, *Tango* was Google's Augmented Reality deprecated project due to its technological issue [58, 28]. Actually, Google's Gemini glitches cost Google 90 billion dollars in stock loss in a single day. Furthermore, *transformer* is a significant natural language process(NLP) model architecture for generative AI. *scalar*, an element used to define a vector space in machine learning, can refer to the number of parameters for an AI model, which is associated with model performance. *Tera* referred to a terabyte, representing a large dataset.

The following top 15 words are the most influential words to increase volatility,

$$\tilde{S_+}^{VOL} = \textit{lancet, forearm, renter, koa, fang, irreversibly, granularity, bully}$$
$$\textit{killing, impropriety, misrepresentation, reimportation, amenable, sayer, spoof}$$

Two themes could be inferred from the extracted words to increase market uncertainty. The first theme could be 'COVID-19' from words, including *lancet* and *forearm*. *lancet* is a major medical journal. *forearm* is a body part relevant to a vaccine. Both could be related to the COVID-19. We secondly could infer 'Innovation and Growth'. The words contained *fang, bully, amenable, killing, granularity, koa*, and *renter*. *fang* referred to FANG, which is an acronym for major technology firms in the US. They were known for their volatile stock prices. Words like *granularity, koa*, and *renter* could be associated with a tech firm's development. The word *renter* indicated the concept of the sharing economy. A representative example of this sharing economy could be cloud service. *koa* represented a web framework for Node.js, which highlighted the back-end technology.

The following words could make an impact on low volatility:

$$\tilde{S_-}^{VOL} = \textit{payday, swine, farming, brod, laryngoscope, moss, socialize}$$
$$\textit{subway, malevolent, entrust, mop, zein, municipalization, awesome, roller}$$

In fact, we could interpret the extracted words in various ways. One theme we could infer was 'Environmental Issues' from words, including *swine, laryngoscope, farming, moss*, and *subway*. *swine* could refer to swine flu. *laryngoscope* could be related to COVID-19. This flu and the pandemic could increase or decrease the volatility of the technology sector. Moreover, words such as *farming, moss*, and *subway* could be associated with modernisation or urbanisation for sustainable development. Or, they could be related to climate change. A few studies supported that environmental issues were correlated with volatility [78, 77].

The Sector Sentiment Model with Only-Risk-Factor (Figure F.1)

The following words were extracted from the $\tilde{p}^{RET}$ model trained with the risk factor section.

$$\tilde{S_+}^{RET} = \textit{wrongfully, mechanic, roe, th, kept, stance, determinable,}$$
$$\textit{escheat, excellence, wake, artificially, magma, wane, ruble, explorer}$$

These words affected the sector sentiment positively. The word set showed a similar theme compared to the previous one (check $\tilde{S_+}^{RET}$ of the 10-K sector model in here). The first theme, 'Innovation and Development,' could be inferred from the words, including *artificially*, *mechanic, excellence*, and *explorer*. *artificially* represents Artificial Intelligence(AI), which is a cornerstone of the fourth industrial revolution. *Explorer* could be interpreted in the same context of *maximizer*, *searcher* in the previous one (check $\tilde{S_+}^{RET}$ of the 10-K sector model in here). The second theme was 'Financial Health'. Words, such as *roe, stance, escheat, wrongfully*, and *ruble* could indicate the theme. *roe* referred to ROE (i.e. Return on Equity). *roe, stance*, and *wrongfully* could symbolise firms' financial condition. *Escheat* could convey asset retention. Also, *ruble* could indicate global market interactions. The ruble is the currency of the Russian Federation.

The following words influenced the sector sentiment negatively.

$$\tilde{S_-}^{RET} = \textit{heat, map, biogen, density, ramification, covid, else, surviving,}$$
$$\textit{pirate, constellation, harmonize, chemotherapy, custody, entrust, kinase}$$

There are a few words that could be interpreted as similar to the 'Technological difficulty' theme mentioned previously (check $\tilde{S_-}^{RET}$ of the 10-K sector model in here). The words included *biogen, surviving, kinase, ramification*, and *chemotherapy*. These words could specifically indicate biotechnology. *biogen, kinase*, and *chemotherapy* are terminology used in biotechnology. *Ramification* and *surviving* could be interpreted as a difficulty to develop biotechnology. Interestingly, this $\tilde{S_-}^{RET}$ model trained with the risk factor section extracted the word *covid* explicitly, which rapidly dropped the sector sentiment with the reduction of stock price.

The following words extracted in the risk factor sector could explain the rise of tech sector uncertainty.

$$\tilde{S_+}^{VOL} = \textit{substituted, seamless, programmatic, reimportation, pressing, prominently clothing, anima, impropriety, spoof, kickback, endocrinologist, educator, polymeric, erroneously}$$

A few words, such as *pressing, prominently*, and *endocrinologist* could show the COVID-19 relevance. Also, *Spoof* appeared again. It could symbolise cybersecurity, including *impropriety, programmatic*, and *erroneously*. Interestingly, *clothing* and *polymeric* were extracted. They could indicate wearable technology in health care.

The following words are the words extracted from the risk factor that contributed to reducing market uncertainty.

$$\tilde{S_-}^{VOL} = \textit{reuse, constriction, knew, mentor, intuit, sodium, sandy, stall, cook, malevolent, snowstorm, maria, residue, mat, swine}$$

5.3.2 Portfolio Level (i.e. Top10 Firms Portfolio)

We generated sentiment scores at a portfolio level. The portfolio consists of the top 10 firms equally listed in the QQQ fund. Top 10 firms denoted the top 10 most invested companies from 1 to 10 in QQQ, as the 50 percentile of the QQQ fund. Note again that the Kalman filter also was applied here.

5.3.2.1 Portfolio Correlation Analysis

The Top10 Sentiment Model with 10-K filing (Figure F.2)

Figure F.2 indicated the predicted sentiment score for the top10 portfolio. These scores were estimated based on the top 10 firms' entire 10-K filing for a given window from 2006 to 2023. The $\tilde{p}^{RET}$ model performed a 76% accuracy rate with 0.22 loss. The previously mentioned sector models(Figure 5.1, FigureF.1) had 78% and 72%, respectively. The $\tilde{p}^{VOL}$ model decreased its accuracy rate to 78% with the 0.20 loss

(5.1: 92%, F.1: 90%), but it still predicted the sector's sentiment labelled with $volatility$ well. The training time was slightly faster, reduced to 13 seconds, as we only used the 10-K filings of the top 10 firms.

Compared to the sector's sentiment score (Figure 5.1), the top10 10-K model (Figure F.2) showed a significantly far clearer trend while noisy signals were removed again through the Kalman filter (the unfiltered model can be found at Appendix F). We found extremely interesting a few phenomena in this model. Firstly, we observed the $\tilde{p}^{RET}$ strongly mirrored the top 10 stock price movement. We could see that the $\tilde{p}^{RET}$ had a steady and moderate increase from 2006 to around 2016. Then, beginning around 2018, $\tilde{p}^{RET}$ showed a rapidly sharp rise, but even after the outbreak of COVID-19. Very similarly, the Top 10 stock rose sharply over a few years in early 2018 and showed a steep drop from 2021 to the beginning of 2023. Then, it increased sharply again. The steep drop from 2021 to 2023 could be stemmed from the COVID-19 pandemic. Several studies argued that quantitative easing(QE) seemed to be significantly and positively related to a higher recovery of the USA's stock market. This could explain the QQQ's sharp increase after the outbreak of COVID-19 [38, 21, 23, 42, 114, 107, 98, 44]. However, the post-COVID-19 surge in sentiment suggested the $\tilde{p}^{RET}$ model did not accurately gauge event significance, a factor likely considered in 10-K filings. This limitation could arise from managers' ignorance about an event's impact at filing time or the model's inability to quantify how much sentiment is driven by the event. Essentially, the model treated significant terms like "COVID-19" or "restrictions" merely as text, without evaluating their actual importance or sentiment contribution.

Secondly, the $\tilde{p}^{VOL}$ model showed interesting points. The $\tilde{p}^{VOL}$ model revealed trends in market volatility, with a sharp increase until around 2010, followed by a decline until around 2016, likely influenced by the 2007-2008 financial crisis and subsequently moderated by quantitative easing (QE) policies. Economists credited the Federal Reserve's fiscal stimulus for mitigating the crisis's effects [38, 21, 106, 42, 23, 114, 75]. A similar volatility movement happened again during the outbreak of COVID-19. Furthermore, this $\tilde{p}^{VOL}$ model indicated a decalcomania-like movement with $\tilde{p}^{LM}$, where increases in volatility seemed to be associated with decreases in $\tilde{p}^{LM}$ and vice versa. Despite this, $\tilde{p}^{LM}$ failed to accurately reflect significant market events, such as the 2008 financial crisis or the COVID-19 pandemic, showing no substantial change in its trend from 2018 to 2024. This lack of responsiveness suggested the $\tilde{p}^{LM}$ model's inability to dynamically mirror the market, likely due to its training dataset being heavily skewed towards negative terms.

Table 5.4 showed the Pearson correlation coefficients of the filtered top10 sentiment model with the entire 10-K filings. All r values indicated a strong correlation with rigorous statistical significance. Compared to Figure 5.1, the top 10 firms revealed clearer correlations by removing the noise associated with the other 90 firms. A strong positive correlation of $r = +.905$ between $\tilde{p}^{RET}$ and $\tilde{p}^{VOL}$ indicated that higher sector positivity was linked with greater uncertainty. Additionally, $r_{\tilde{p}^{RET},Stock} = +.956$ showed that increases in sector stock prices correlated with more positive sentiment, and similarly, a strong positive correlation existed between $\tilde{p}^{VOL}$ and stock prices, suggesting that stock prices rose with increased market uncertainty. Strong negative correlations of $r_{\tilde{p}^{RET},\tilde{p}^{LM}} = -.922$ and $r_{\tilde{p}^{VOL},\tilde{p}^{LM}} = -.903$ indicated that decreases in

$\tilde{p}^{LM}$ were associated with the increased market sentiment and uncertainty, and vice versa. Also, $\tilde{p}^{LM}$ had a strong negative correlation with *Stock*. The $\tilde{p}^{LM}$ model was biased toward having a negative tone.

The Top10 Sentiment Model with Only-Risk-Factor (Figure F.3)

Figure F.3 showed the top 10 firms' estimated sentiment scores by training only the risk factor section. Compared to the 10-K top10 model (Figure F.2), this model(Figure F.3) performance was decreased slightly to 71% with the 0.23 loss ($\tilde{p}^{RET}$) and 74% with the 0.21 loss ($\tilde{p}^{VOL}$). This could be because of the smaller size of the vocabulary of the risk factor than the entire 10-K filings. Then, the training time also was reduced to 8 seconds. Due to the same reason, we observed that the $\tilde{p}^{RET}$ model steadily increased. However, this model did not reflect the market situation well. This model gradually increased during the technology industry boom time and very slightly decreased after the outbreak of COVID-19. Also, the $\tilde{p}^{VOL}$model showed the market uncertainty steadily fell during the same period. The $\tilde{p}^{RET}$ model dynamically, albeit minimally, reflected market changes, while the $\tilde{p}^{VOL}$ model struggled due to its focus on the risk factor section dominated by negative tones. When using the model trained on risk factors, we must remember it might offer a pessimistic market view. Furthermore, the $\tilde{p}^{LM}$ model indicated a slow, moderate shift towards negativity, showing it lacks dynamic market responsiveness, similar to patterns seen in the earlier $\tilde{p}^{LM}$ model(Figure F.2). This negative bias distorted the analysis of the portfolio level.

When examining the Pearson correlation Table F.3, the correlations including $\tilde{p}^{VOL}$ showed differently compared to the 10-K top 10 sector model(Table F.2). We observed that all correlation, such as $r_{\tilde{p}^{VOL},\tilde{p}^{RET}}$(=+.145), $r_{\tilde{p}^{VOL},\tilde{p}^{LM}}$(=-.349), and $r_{\tilde{p}^{VOL},Stock}$(=+.185), became to have a weak linear correlation. The $r_{\tilde{p}^{VOL},\tilde{p}^{RET}}$ and the $r_{\tilde{p}^{VOL},Stock}$ even were close to zero, implying that they lost their correlation. We could assume that the $\tilde{p}^{VOL}$ score was not calculated correctly in the model trained with the risk factor section.

5.3.2.2 Portfolio Most Influential Words

The Top10 Sentiment Model with 10-K filing (Figure F.2)

The following set of words positively affected the portfolio return. It was trained with the entire 10-K filings.

$$\tilde{S_+}^{RET} = \textit{battery, solar, revolving, screen, musk, commensurate, secondary,}$$
$$\textit{roadster, separately, dealer, motor, importation, installation, misstatement, convert}$$

The word sets extracted from the top10 10-K model seemed to indicate the 'Electric cars'. The 10-K model includes words like *roadster, musk, supercharger, cluster, roof, sedan,* and *dangerous*. These words explicitly indicated 'Electric cars'. Notably, the same theme commonly appeared in the following top10 risk-factor model. From the observation, we could infer that 'Electric cars' have a strong correlation with the technology sector return.

The following set of words, extracted from the entire 10-K, negatively impacted the portfolio return.

$$\tilde{S_-}^{RET} = \textit{formation, wholly, family, entire, incorporated, dependence, nationwide, misappropriate, wrong, sign, carrier, treat, subcontractor, entrust, remotely}$$

The 10-K word set to affect return negatively seemed to reveal generic words, hindering catching a theme. However, the following $\tilde{S_-}^{RET}$ in the top10 risk factor model seemed to highlight a distinct theme.

The following words set increased or decreased the portfolio uncertainty.

$$\tilde{S_+}^{VOL} = \textit{autonomous, ford, draft, berlin, thin, neural, audible, accomplished, log, understatement, identifier, segregation, subjectivity, race, pilot}$$
$$\tilde{S_-}^{VOL} = \textit{restrain, club, deflation, voucher, ugh, explorer, insecure, ticket, tester, murphy, fifty, seminar, resilience, hospital, mary}$$

The positive word set had a common theme, 'Electric cars', with $\tilde{S_+}^{RET}$ in the top10 10-K model. The words indicating the topic include *autonomous, neural, pilot ford, berlin,* and *race*. *autonomous, neural, and pilot* could refer to self-driving, while *ford, berline, and race* could represent an automobile industry. From this observation, 'Electric cars' also was the topic to increase the technology market's uncertainty as well as the market's positivity.

The words from $\tilde{S_-}^{VOL}$ such as *restrain, resilience, tester,* and *insecure* could indicate a topic, 'Stability and Resilience". The continuous efforts(*tester*) of a tech firm to overcome technological difficulties(*insecure, restrain*) could improve firm stock's stability(*resilience*), contributing to lower volatility.

The Top10 Sentiment Model with Only-Risk-Factor (Figure F.3)

The following word sets to impact positively on return are

$$\tilde{S_+}^{RET} = \textit{roadster, misconduct, musk, supercharger, agricultural, detriment, cluster, owing, observe, rigor, roof, sedan, visible, ramping, dangerous}$$

From the words such as *battery, solar, musk, screen, roadster*, and *motor*. This set could indicate 'Electric Cars'. Interestingly, this topic appeared in the top10 10-K model to

increase return and volatility, suggesting investors should carefully pay attention to the electric car industry. It could be the next alpha(α) signal to beat the market.

The following word set seemed to indicate a more specific theme compared to $\tilde{S_-}^{RET}$ in the top10 10-K model.

$$\tilde{S_-}^{RET} = \textit{session, identifier, vista, portable, dram, snap, wrong, snow, node, attendant, printed, peripheral, palm, composed, pride}$$

The theme "Supply chain risk", for instance, could be revealed from words like *subcontractor, entrust, dependence*, and *nationwide*. Several studies showed the risk of supply chains for high-technology industries severely increased due to geopolitical disruptions [111, 112, 70, 120]. *subcontractor, entrust, dependence* could indicate firms' dependency(*dependence*) on their productions and services. *nationwide* could represent supply chain risk occurring globally

The following word set extracted from the risk factor increased or decreased the portfolio uncertainty.

$$\tilde{S_+}^{VOL} = \textit{resale, eligible, tender, generating, according, circuit, virtual, duplicate, franchise, deposit, air, though, representation, transact, attorney}$$
$$\tilde{S_-}^{VOL} = \textit{depot, host, sea, club, array, essential, membership, accountability, subjective, surface, subscriber, azure, evidence, care, bing}$$

The $\tilde{S_+}^{VOL}$could indicate a theme, 'Innovation and development' to increase market uncertainty. *circuit, generating* could symbolise generative AI, and its semiconductor chip, GPU. *virtual* could refer to the metaverse. *transact* could indicate transaction technology like blockchain or digital payment systems. Several studies found that firms' RD investment in these fields, being represented by innovation, has a strong positive relationship with volatility [80, 40, 47]. The topic, 'Innovation', from the model supported the argument of these studies.

The $\tilde{S_-}^{VOL}$could refer to a theme, 'Infrastructure and Services', to decrease market uncertainty. *depot, host* could indicate a data centre and its hosting service. Also, *sea* could be related to undersea cables for data transmission. *club, membership* can represent a social infrastructure(or network) for tech operations. Several studies found digital and traditional infrastructure investments are key components to stabilising and driving forward economic growth as well as reducing market uncertainty[14, 11, 45]. These studies support the model's argument that 'Infrastructure and Services' could decrease market uncertainty.

5.3.3 Company Level (i.e. Nvidia)

5.3.3.1 Nvidia Correlation Analysis

Nvidia Sentiment Model with 10-K filings (Figure F.4)

Figure F.4 represented Nvidia's predicted sentiment scores trained with the entire 10-K filing. The $\hat{p}^{RET}$ model performed a 75% with 0.19 loss and the $\hat{p}^{VOL}$ showed a 69% accuracy with 0.25 loss. Our sentiment prediction model predicted a single firm's sentiment fairly well. It took 7 seconds.

The $\hat{p}^{RET}$ model seemed to reveal a periodic pattern of the sentiment tone. Its sentiment tone dramatically became positive from 2006 to 2011 and vastly became negative from 2014 to 2020. In the same period, Nvidia's stock price remained the same. Then, it sharply and rapidly became positive. The stock was hugely impacted by the outbreak of COVID-19 and rebounded due to the US government's aggressive QE policy and Nvidia's dominant position in the GPU market in the fourth industrial revolution. We assumed that Nvidia's periodic rise and fall in their sentiment might have stemmed from the firm's internal optimism and the market not responding to that optimism. Furthermore, the $\hat{p}^{VOL}$ model showed a similar movement to the $\hat{p}^{RET}$ model. This model could capture the aftermath of COVID-19, but it is not that great. As our model was not trained with 2024 data, the $\hat{p}^{VOL}$ could not capture the current soaring as well as $\hat{p}^{RET}$, and $\hat{p}^{LM}$. The $\hat{p}^{LM}$, interestingly, showed a decalcomania-like pattern to the $\hat{p}^{RET}$. Also, they could capture COVID-19's aftermath.

Table F.4 revealed all r values(i.e. all correlations) showed the same movement compared to the 10-K top10 model's correlation (Table F.2), but represented a lower correlation power. For instance, Nvidia showed a moderate positive correlation at $r_{\tilde{p}^{RET},\tilde{p}^{VOL}}$(=+.612), whereas the 10-K top10 model (Table F.2) showed a strong positive correlation at $r_{\tilde{p}^{RET},\tilde{p}^{VOL}}$(=+.905).

Nvidia Sentiment Model with Only-Risk-Factor (Figure F.5)

Figure F.5 indicated Nvidia sentiment prediction score trained with the risk factor. Compared to the 10-K Nvidia sentiment model (Figure F.4), the $\hat{p}^{RET}$ showed almost the same performance, and the $\hat{p}^{VOL}$ increased 6% accuracy. The prediction took less time. It took 4 seconds as fewer tokens were used(i.e. 2,201). In the same comparison between Figure F.5 and Figure F.4, the $\hat{p}^{RET}$ and the $\hat{p}^{VOL}$ showed a similar movement, whereas the risk factor $\hat{p}^{LM}$model showed a gradually falling movement. We assumed this was also because the $\hat{p}^{LM}$ model dataset was predominantly filled with negative words, as previously mentioned in Figure F.2.

Table F.5 revealed a meaningful correlation compared to the 10-K Nvidia model (Table F.4). The risk-factor model (Table 5.5) showed a slightly stronger correlation but with more rigorous statistical significance in both $r_{\hat{p}^{VOL},\hat{p}^{RET}}$ and $r_{\hat{p}^{VOL},Stock}$. The correlation results of the 10-K Nvidia model were not actually reliable as almost all correlations did not show statistical significance, except for the RET and VOL pair. For your consideration, both $\hat{p}^{LM}$'s correlations were less informative for our project. Even $\hat{p}^{LM}$'s movements inaccurately mirrored Nvidia stock's behaviour.

5.3.3.2 Nvidia Most Influential Words

Nvidia Sentiment Model with 10-K filing (Figure F.4)

The following words were the most impactful words in Nvidia's $\hat{p}^{RET}$ sentiment score prediction, positively or negatively, respectively.

$$\tilde{S_+}^{RET} = \textit{chain, weak, tender, russia, talent, ninth, thermal, conflict, dispose, exclusion, strict, failing, connect, governmental, favor}$$
$$\tilde{S_-}^{RET} = \textit{redemption, grid, initiative, petition, eastern, big, retrospective, branded, revolving, enactment, reducing, drone, joint, pursuit, broadcast}$$

The word set to impact positively on Nvidia sentiment could be interpreted into two categories. The first category could be 'Innovation and Development'. This theme appeared again in the sector analysis. Words such as *talent, thermal, failing, strict*, and *connect* could be related to a firm's innovation. *Talent* could indicate that talented employees were drivers of innovation. *Thermal* could convey innovative GPU development, which is Nvidia's main product, as temperature control is required for chip's higher performance [88]. Also, words, including *chain, favour, exclusion*, and *tender* could symbolise the second theme, 'GPU market domination'. *tender* could represent the tendering process, which is a key component of the supply chain management system. We could interpret its GPU market domination as being caused by its innovation and strategic supply chain management, including tender processing, leading to the exclusive market position [69, 3].

The word set to influence the return negatively could be interpreted in a various way. ChatGPT4([84]) suggested that 'Market Challenges' could have a negative impact on Nvidia's sentiment. The relevant words for the theme included *redemption, grid, initiative, petition, eastern, big, retrospective, branded, revolving, enactment, reducing, drone, joint,* and *pursuit*. ChatGPT4([84]) argued that they reflected various operational and market challenges. For Nvidia, this could entail navigating energy regulations (*grid*), responding to legal and regulatory *petitions*, while managing its *brand* in a competitive market *big, eastern* by adapting to laws affecting its business model *initiative* or product offerings (*enactment*).

The following extracted words mostly affected the rise of Nvidia's volatility.

$$\tilde{S_+}^{VOL} = \textit{tracing, starting, white, attempt, derive, turn, antitrust, pace, retroactively, popularity, severe, subjectivity, accessible, percent, travel}$$

We could interpret "Nvidia, Venture Capitalist" from words such as *starting, attempt*, and *derive*. Nvidia currently is ramping up its venture investment [2, 4, 89]. Venture

investment in a startup (*starting*) that *attempts* to *derive* innovation could be risky and subsequently could increase volatility.

The following words mostly influenced the reduction of Nvidia's market uncertainty.

$$\tilde{S_-}^{VOL} = \textit{settle, dow, onto, occurrence, eventually, strike, unfair, allegation, returned, grown, virus, programmer, fault, near}$$

The words, such as *virus, grown, returned, settle*, and *near, occurrence*, could refer to "Stability from Alleviating the aftermath of COVID-19". They may hint at recovery(*settle, returned, near, occurrence*) from setbacks(*virus, grown*), signalling the reduction of volatility.

Nvidia Sentiment Model with Only-Risk-Factor (Figure F.5)

The following positive word set seemed to show more accurate information to identify a positive topic on return.

$$\tilde{S_+}^{RET} = \textit{chain, conflict, mineral, group, implementation, card, item, weak, antitrust, original, intangible, importance, pace, pose, thermal}$$

For example, the word set, including *card, mineral, original, thermal, chain*, and *intangible*, could symbolise 'Graphics Processing Unit(GPU)', which has been the cash cow for Nvidia. Those words seemed to directly be indicative of GPU. *card* could represent a graphic card. *mineral* could indicate core raw materials for GPU manufacturing [32]. *chain* could refer to the supply chain of Nvidia that had a dominant position in the GPU market. *thermal* could indicate the innovative development of GPU. *original* and *intangible* could indicate Nvidia GPU's originality and its intellectual property.

The following words were negatively influential on Nvidia's return sentiment.

$$\tilde{S_-}^{RET} = \textit{directive, console, video, suit, three, go, pursuit, innovative, interested, floating, initiative, discovered, elsewhere, generating, member}$$

Note that the words were extracted from the risk factor section where negative words were prevalent. This section contained valuable information to analyse firms' risks. So, we assume that the risk section could offer better informative insight for analysing their negative sentiment. From words like *console* and *video*, we could deduce the 'Video Game Industry' topic and its negative impact on return. Actual historical revenue data of Nvidia supported our interpretation. The gaming segment's revenue has been reduced [83].

The following risk-factor-centred word set could offer informative ideas for the market uncertainty rise.

$$\tilde{S_+}^{VOL} = \textit{responsible, procedure, social, restricted, included, severe, forecasting, seeking, distribute, providing, full, begin, pose, mitigate, depending}$$

The theme "Regulation and social responsibility on AI" emerged from words, including *social, responsible, severe, restricted*, and *procedure*. Given that the Environmental, Social, and Governance (ESG) ratings effectively influence return and volatility, ESG scores become important financial indicators [68, 16, 41]. Consequently, we could interpret from the word set that the importance(*severe*) of self-regulation(*restricted*) for its product manufacturing *procedure* as part of their social responsibility (*social*, and *responsible*) on AI.

The following word set could refer to the 'External risks' topic.

$$\tilde{S_-}^{VOL} = \textit{shareholder, hacker, fault, near, architecture, exercise, dilute, occurrence, programmer, cessation, worm, geographic, miss, fourth, confidential}$$

The words relevant to the 'External risks' theme included *worm, hacker, fault, occurrence, confidential*, and *geographic*. Cyber security, as an external risk, could be deduced from words like *worm, hacker, fault, occurrence, and confidential*. Also. *geographic* could indicate a geopolitical issue as the US government is restricting cutting-edge AI chip export to China [97]. While cyber security or geopolitical issues could heighten volatility, we could interpret that Nvidia's effective risk management strategies, in the context of low volatility, mitigated its volatility. However, note that the model could be biased as managers' biases were reflected in the training data, a risk factor. Just because managers' arguments were positive to those issues, that does not mean risks did not disappear[123].

Chapter 6

Conclusions

6.1 Conclusions

Our suggested models estimated sentiment scores at three stakeholder levels: sector level, portfolio level, and company level. Also, each model was trained with either the entire 10-K fillings or the risk factor section.

At the sector level, the sector model trained with the entire 10-K provided more informative sentiments than the risk-factor-centred sector model. The 10-K model outperformed the risk factor model for both the return-labelled sentiment(i.e. $\tilde{p}^{RET}$) and volatility-labelled sentiment(i.e. $\tilde{p}^{VOL}$). The accuracies for $\tilde{p}^{RET}$ or $\tilde{p}^{VOL}$ were higher by 5% and 2% respectively. The 10-K model took more time for training though. Notably, $\tilde{p}^{VOL}$, in the 10-K sector model, showed periodic fluctuation trends in volatility, whereas $\tilde{p}^{RET}$ was less fluctuated. In terms of correlation analysis, the 10-K sector showed a stronger correlation than the risk-factor sector model. For the qualitative analysis with the most impactful words, both models seemed to exhibit similar topics. The baseline model (i.e. $\tilde{p}^{LM}$) did show a distorted performance compared to both the 10-K sector model and the risk factor sector model. This was because the baseline model was biased to be negative as the training set was predominately filled with negative tokens. It is noteworthy that the Kalman filter should be applied for sector analysis to control noise.

At the portfolio level, the model based on the entire 10-K generally revealed more informative sentiments than the model with the risk factor. The 10-K portfolio model outperformed the risk factor portfolio model for both $\tilde{p}^{RET}$ and $\tilde{p}^{VOL}$. The accuracies for $\tilde{p}^{RET}$ or $\tilde{p}^{VOL}$ were higher by 5% and 4% respectively. The training time took longer than the risk-factor model. When it comes to correlation analysis, the 10-K sector depicted a way stronger correlation for all pairs. In the 10-K portfolio model, $\tilde{p}^{RET}$ strongly mirrored the top 10 portfolio stock price movement. $\tilde{p}^{RET}$ seemed to show no significant response to macroeconomic external factors such as the 2008 financial crisis and COVID-19. On the other hand, $\tilde{p}^{VOL}$ revealed trends in volatility response to the external factors. For the influential word analysis, the risk factor portfolio model seemed to highlight a distinct topic. Thus, we recommended employing two models(i.e. both the 10-K one and the risk factor one) for a portfolio-level analysis. Again, it is

remarkable that the Kalman filter was also required for a portfolio analysis.

At the company level, the company-specific model using risk factors yielded sentiments with greater informational value than the model focusing on the complete 10-K documents. The risk factor model outperformed the 10-K-centred model for both $\hat{p}^{RET}$ and $\hat{p}^{VOL}$. The accuracy levels for $\hat{p}^{RET}$ and $\hat{p}^{VOL}$ saw increases of 1% and 6%, respectively. The training time was shorter. For both the risk-factor company model and the 10-K company model, $\tilde{p}^{RET}$ and $\tilde{p}^{VOL}$ exhibited a similar movement. For correlation analysis, the model centred on risk factors demonstrated a marginally higher correlation, accompanied by more robust statistical significance. In the most influential word analysis, both models provided a distinct and detailed word, leading to capturing the theme. Hence, at the company level, we advised using the risk-factor-centred model for trend and correlation analysis while suggesting the use of both models for word analysis.

6.2 Limitations and Future Works

Our suggested models can not currently capture meaningful phrase words as the model essentially calculates sentiment-charged words based on a bag of words. For instance, the model will extract the phrase word 'chef executive officers(CEO)' in a word base separately and then evaluate whether each word can be sentiment-charged words concerning the dependent variables. In this word-based separation process, it loses the original meaning, which is CEO. Hence, in the future, we suggest to add a function that contains contextual information on the model. bi-gram or n-gram can be examples.

Industrial sentiment score prediction does not consider the allocation proportion of the QQQ portfolio. In the QQQ ETF fund, the top 10 firms take around 45% allocation proportion of the total in 2023, and the rest of 90 firms take the rest of 55% proportion. The portfolio is reconstructed annually. So, to predict a robust industrial-level sentiment score, the scores should consider the portfolio allocation proportion. In our industrial sentiment trending analysis, however, all sentiment scores are equally considered as the portfolio rebalancing data is restricted to attain. For instance, the return-labelled sentiment score of Apple Inc. in 2023 is 0.006, whereas Amazon.com Inc.'s sentiment in 2023 is -0.009. We used these scores without weighting their portfolio proportion. In 2023, the QQQ allocated Apple at 9.22% and Amazon at 4.83%. Thus, the weighted sentiment scores(i.e. 0.05532 = 0.006 * 9.22 for Apple and -0.04347 = -0.009 * 4.83 for Amazon) are required for a robust sentiment score calculation. In the future, we suggest that the sentiment score at the date should consider the allocation weight of the portfolio of the same date.

Our model can not adapt to the latest firms' return or volatility because our model calculation only works with the filings released at the publication date. In other words, the predicted sentiment score is an annual data point at which the filing is released. If we have more latest textual information to represent a firm, our model would generate more recent sentiment scores. We suggest using 10-Q filling, which is a comprehensive report of a firm like 10-K but must be submitted quarterly.

Bibliography

[1] Securities and exchange commission final rule, release no. 33–8591 (fr-75). http://sec.gov/rules/final/33-8591.pdf, 2005.

[2] Nvidia for startups: Venture capital. https://www.nvidia.com/en-gb/startups/venture-capital/, 2024. Accessed: 2024-03-27.

[3] Nvidia industries: Retail - supply chain management. https://www.nvidia.com/en-gb/industries/retail/supply-chain-management/, 2024. Accessed: 2024-03-27.

[4] Nvidia investments. https://blogs.nvidia.com/blog/nvidia-investments/, 2024. Accessed: 2024-03-27.

[5] Invesco 2023. About invesco qqq etf. Accessed: 2024-03-21.

[6] Sassan Alizadeh, Michael W. Brandt, and Francis X. Diebold. Range-based estimation of stochastic volatility models. *The Journal of Finance*, 57(3):1047–1091, 2002.

[7] Analytics Vidhya. Stemming vs lemmatization in nlp: Must know differences, 2022. Accessed: 22-03-2024.

[8] American Statistical Association. American statistical association releases statement on statistical significance and p-values, 2016. For more information: Ron Wasserstein, ron@amstat.org.

[9] Sharad Asthana and Steven Balsam. The effect of edgar on the market reaction to 10-k filings. *Journal of Accounting and Public Policy*, 20(4-5):349–372, 2001.

[10] D. J. Benjamin, J. Berger, M. Johannesson, B. A. Nosek, E. Wagenmakers, R. Berk, et al. Redefine statistical significance. Jul 2017. https://doi.org/10.31234/osf.io/mky9j.

[11] Frédéric Blanc-Brude, Wilhelm Schmundt, Thomas Bumberger, Roman Friedrich, Bernhard Georgii, Abhishek Gupta, Leonard Lum, and Maikel Wilms. Infrastructure strategy 2022: A pivot to the digital frontier. Boston Consulting Group, 3 2022. Article.

[12] Sander Blomme and Julie Dedeyne. Predicting the effect of 10-k, 10-q and 8-k company reports on abnormal stock returns using finbert nlp methods. Technical report, University of Ghent, 2020.

[13] Svetlana Borovkova and Philipp Lammers. Sector news sentiment indices. Available at SSRN 3080318, 2017.

[14] Marcel Brinkman and Vijay Sarma. Infrastructure investing will never be the same. McKinsey & Company, 8 2022. Article.

[15] J. L. Campbell, H. Chen, D.S. Dhaliwal, H. Lu, and L.B. Steele. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455, 2014.

[16] Gunther Capelle-Blancard and Anastasia Petit. Every little helps? esg news and stock market reaction. *Journal of Business Ethics*, 157:543–565, 2019.

[17] Luke Chan and Nicolás Devia-Valbuena. In the global rush for lithium, bolivia is at a crossroads, December 2023. Accessed: 2024-03-26.

[18] S. Chang. Risk factor disclosures and ceo overconfidence. College of Business Administration Seoul National University, 2019.

[19] S. Che et al. Anticipating corporate financial performance from ceo letters utilizing sentiment analysis. *Mathematical Problems in Engineering*, 2020, 2020.

[20] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies (RFS)*, Forthcoming. Available at SSRN: https://ssrn.com/abstract=1807265 or http://dx.doi.org/10.2139/ssrn.1807265.

[21] Han Chen, Vasco Cúrdia, and Andrea Ferrero. The macroeconomic effects of large-scale asset purchase programs. Staff Report 527, Federal Reserve Bank of New York, December 2011.

[22] Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

[23] Vasco Cúrdia. How stimulatory are large-scale asset purchases? *FRBSF Economic Letter*, 08 2013.

[24] Angela K. Davis, Weili Ge, Dawn Matsumoto, and Jenny Li Zhang. The effect of manager-specific optimism on the tone of earnings conference calls, January 13 2014. CAAA Annual Conference 2012, Available at SSRN: https://ssrn.com/abstract=1982259 or http://dx.doi.org/10.2139/ssrn.1982259.

[25] J.E. Dobson. On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities*, 5:431–449, 2023.

[26] J.S. Doran, D.R. Peterson, and S.M. Price. Earnings conference call content and stock price: The case of reits. *Journal of Real Estate Finance and Economics*, 45(2):402–434, 2010.

[27] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*, volume 38. OUP Oxford, 2012.

[28] Jillian D'Onfro. Google is going to shut down its once-hyped augmented reality project, tango, December 2017. Accessed: 2024-03-26.

[29] Minitab Blog Editor. How to correctly interpret p values, 2014. Topics: Hypothesis Testing.

[30] J. Engelberg. Costly information processing: Evidence from information announcements. In *AFA 2009 San Francisco Meetings Paper*, 2008. Available at SSRN: http://ssrn.com/abstract=1107998.

[31] J. Engelberg, A.V. Reed, and M.C. Ringgenber. How are shorts informed? short sellers, news, and information processing. *Journal of Financial Economics*, 105(2):260–278, 2012.

[32] Euromines. Key value chain electronics euromines, 2020.

[33] Nicky J. Ferguson, Dennis Philip, Herbert Lam, and Jie Michael Guo. Media content and stock returns: The predictive power of press, May 27 2015. Available at SSRN: https://ssrn.com/abstract=2611046.

[34] S.P. Ferris, G.Q. Hao, and M. Liao. The effect of issuer conservatism on ipo pricing and performance. *Review of Finance*, 7(3):993–1027, 2013.

[35] T. Fields, T. Lys, and L. Vincent. Empirical research on accounting choice. *Journal of Accounting and Economics*, 31:255–307, 2001.

[36] Joshua J. Filzen. The information content of risk factor disclosures in quarterly reports. *Accounting Horizons*, 29(4):887–916, Dec 2015.

[37] Jim Frost. Why are p values misinterpreted so frequently?, 2014.

[38] Joseph Gagnon, Matthew Raskin, Julie Remache, and Brian Sack. Large-scale asset purchases by the federal reserve: Did they work? Staff Report 441, Federal Reserve Bank of New York, March 2010.

[39] Diego Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[40] Sami Gharbi, Jean-Michel Sahut, and Frédéric Teulon. R&d investments and high-tech firms' stock return volatility. *Technological Forecasting and Social Change*, 88:306–312, 2014.

[41] Guido Giese, Linda-Eling Lee, Dimitris Melas, Zoltán Nagy, and Laura Nishikawa. Foundations of esg investing: How esg affects equity valuation, risk, and performance. *The Journal of Portfolio Management*, 45:69–83, 06 2019.

[42] S. Gilchrist and E. Zakrajšek. The impact of the federal reserve's large-scale asset purchase programs on corporate credit risk. *Journal of Money, Credit and Banking*, 45:29–57, 2013.

[43] Alexander Glodd and Diana Hristova. Extraction of forward-looking financial information for stock price prediction from annual reports using nlp techniques.

In *Proceedings of the 56th Hawaii International Conference on System Sciences*, page 10, 2023. Rights: Attribution-NonCommercial-NoDerivatives 4.0 International.

[44] Pierre-Olivier Gourinchas, ebnem Kalemli-Özcan, Veronika Penciakova, and Nick Sander. Fiscal policy in the age of covid: Does it 'get in all of the cracks?'. Working Paper 29293, National Bureau of Economic Research, September 2021.

[45] Lester Gunnion. Infrastructure investment: An economist's view from the ground up, 7 2021. Article.

[46] Y. Guo and L. Zhou. Textual tone in corporate financial disclosures: a survey of the literature. *International Journal of Disclosure and Governance*, 17:101–110, 2020. Received October 25, 2019; Published June 27, 2020; Issue Date September 1, 2020.

[47] B. Hai, Q. Gao, X. Yin, and J. Chen. R&d volatility and market value: the role of executive overconfidence. *Chinese Management Studies*, 14(2):411–431, 2020.

[48] Elaine Henry and Andrew J. Leone. Measuring qualitative information in capital markets research, April 2009. Available at SSRN: https://ssrn.com/abstract=1470807 or http://dx.doi.org/10.2139/ssrn.1470807.

[49] Jörg Hering. The annual report algorithm: Retrieval of financial statements and extraction of textual information. Friedrich-Alexander-Universität Erlangen-Nürnberg, 11 2016. First Version: October 4, 2016. Current Version: November 28, 2016.

[50] O. Hope, D. Hu, and H. Lu. The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, 2016. forthcoming.

[51] Allen H. Huang, Amy Zang, and Rong Zheng. Evidence on the information content of text in analyst reports. *Accounting Review*, April 2014. Forthcoming.

[52] Xuan Huang, Siew Hong Teoh, and Yinglei Zhang. Tone management, August 21 2013. The Accounting Review, Forthcoming, Available at SSRN: https://ssrn.com/abstract=1960376 or http://dx.doi.org/10.2139/ssrn.1960376.

[53] IBM. Deep learning. https://www.ibm.com/topics/deep-learning. Accessed: 2024-04-01.

[54] R.D. Israelsen. Tell it like it is: disclosed risks and factor portfolios. Working paper, 2014.

[55] Narasimhan Jegadeesh and Di Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110:712–729, 2013.

[56] R. Johnson and T. Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, volume 1, pages 562–570, 2017.

[57] Kasper Regenburg Jønsson and Jonas Burup Jako. Predicting stock performance using 10-k filings: A natural language processing approach employing con-

volutional neural networks. Technical report, Copenhagen Business School, 2020.

[58] Jacob Kastrenakes. Google's project tango is shutting down because arcore is already here, December 2017. Accessed: 2024-03-26.

[59] 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.

[60] Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

[61] C. Kim, K. Wang, and L. Zhang. Readability of 10-k reports and stock price crash risk. *Contemporary Accounting Research*, 36:1184–1216, 2019.

[62] Y. Kim. Convolutional neural networks for sentence classification. http://arxiv.org/abs/1408.5882, 2014.

[63] D. Kingsley, M. Solomon, and K. Jaconi. Sec risk factor disclosure rules. https://corpgov.law.harvard.edu/2021/12/22/sec-risk-factor-disclosure-rules/, 2021.

[64] Ralph S.J. Koijen, Tomas J. Philipson, and Harald Uhlig. Financial health economics. *Econometrica*, 84:195–242, 2016.

[65] S. P. Kothari, X. Li, and J. Short. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. *The Accounting Review*, 84:1639–1670, 2009.

[66] S. P. Kothari, S. Shu, and P. Wysocki. Do managers withhold bad news? *Journal of Accounting Research*, 47:241–276, 2009.

[67] T. Kravet and V. Muslu. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, 18(4):1088–1122, 2013.

[68] Mario La Torre, Fabiomassimo Mango, Arturo Cafaro, and Sabrina Leo. Does the esg index affect stock return? evidence from the eurostoxx50. *Sustainability*, 12(16):6387, 2020.

[69] Kif Leswing. Meet the $10,000 nvidia chip powering the race for a.i. https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html, 2023. Accessed: 2024-03-27.

[70] J. Lewis. *Learning the superior techniques of the barbarians China's pursuit of semiconductor independence*. CSIS, Washington, 2019.

[71] F. LI. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

[72] Nan Li, Xun Liang, Xinli Li, Chao Wang, and Desheng Dash Wu. Network environment and financial risk using machine learning and sentiment analysis. *Human and Ecological Risk Assessment: An International Journal*, 15(2):227–252, 2009.

[73] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[74] Tim Loughran and Bill McDonald. Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics*, 109:307–326, 2013.

[75] Stephan Luck and Thomas Zimmermann. Ten years later—did qe work?, May 2019.

[76] CFO Magazine. Sec pushes companies for more risk information. August 2, 2010, 2010.

[77] M.T. Majeed, M. Mazhar, and S. Sabir. Environmental quality and output volatility: the case of south asian economies. *Environmental Science and Pollution Research*, 28:31276–31288, 2021.

[78] Muhammad Majeed and Maria Mazhar. Environmental degradation and output volatility: A global perspective. *Pakistan Journal of Commerce and Social Science*, 13:180–208, 03 2019.

[79] Junaid Maqbool, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, and Ishfaq Ali Ganaie. Stock prediction by integrating sentiment scores of financial news and mlp-regressor: A machine learning approach. *Procedia Computer Science*, 218:1067–1078, 2023.

[80] M. Mazzucato and M. Tancioni. R&d, patents and stock return volatility. *J Evol Econ*, 22:811–832, 2012.

[81] K. Mishev, A. Gjorgjevikj, I. Vodenska, L.T. Chitkushev, and D. Trajanov. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131662–131682, 2020.

[82] Newcastle University. Critical region and confidence interval. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/hypothesis-testing/critical-region-and-confidence-interval.html, 2024. Accessed: 24-03-2024.

[83] Nvidia. 2023 annual report, 2023.

[84] OpenAI. Chatgpt (4) [large language model]. https://chat.openai.com, 2024.

[85] Jangsi Park. Don't be overwhelmed by p-value, 2016.

[86] Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011.

[87] H. M. Peirce. Sec harming investors and helping hackers: Statement on cybersecurity risk management, strategy, governance, and incident disclosure, 2023.

[88] Alok Prakash, Hussam Amrouch, Muhammad Shafique, Tulika Mitra, and Jörg Henkel. Improving mobile gaming performance through cooperative cpu-gpu thermal management. In *Proceedings of the 53rd Annual Design Automation Conference*, DAC '16, New York, NY, USA, 2016. Association for Computing Machinery.

[89] PYMNTS. Nvidia invests in 35 ai companies in 2023. https://www.pymnts.com/artificial-intelligence-2/2023/nvidia-invests-in-35-ai-companies-in-2023/, December 2023. Accessed: 2024-03-27.

[90] R. Ren, D. D. Wu, and T. Liu. Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, 13(1):760–770, Mar 2019.

[91] Reuters. 'refco risks boiler-plate disclosure.' by scott malone. http://w4.stern.nyu.edu/news/news.cfm?doc_id=5094, 2005.

[92] M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik, and D. Trajanov. Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access*, 12:7170–7198, 2024.

[93] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems*, 27(2):12, Mar 2009.

[94] T. Scott. Incentives and disincentives for financial disclosure: Voluntary disclosure of defined benefit pension plan information by canadian firms. *The Accounting Review*, 69:26–43, 1994.

[95] SEC. Form 10-k instructions. http://www.sec.gov/about/forms/form10-k.pdf, 2010.

[96] Securities and Exchange Commission. Form 10-k. https://www.sec.gov/files/form10-k.pdf, 2024. Annual Report Pursuant to Section 13 or 15(d) of The Securities Exchange Act of 1934, MB Number: 3235-0063, Expires: December 31, 2026, Estimated average burden hours per response: 2,249.36.

[97] Demetri Sevastopulo and Qianer Liu. Us-china doomsday scenario not likely to happen, says nvidia's jensen huang. Financial Times, 10 2023.

[98] Ünal Seven and Fatih Yılmaz. World equity markets and covid-19: Immediate response and recovery prospects. *Research in International Business and Finance*, 56:101349, 2021.

[99] Jun Sha. A data pipeline framework for automated extraction of risk factors from sec-10k filings, 2023.

[100] Dev Shah, Haruna Isah, and Farhana Zulkernine. Predicting the effects of news sentiments on the stock market. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4705–4708. IEEE, 2018.

[101] S.M. Shuhidan, S.R. Hamidi, S. Kazemian, S.M. Shuhidan, and M.A. Ismail. Sentiment analysis for financial news headlines using machine learning algorithm.

In *Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018. KEER 2018. Advances in Intelligent Systems and Computing*, volume 739, Singapore, 2018. Springer.

[102] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *Proc. Int. Workshop Hum.-Comput. Interact. Knowl. Discovery Complex Unstructured Big Data*, pages 77–88, 2013.

[103] S. Sohangir, D. Wang, A. Pomeranets, et al. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(3), 2018.

[104] V. Song, H. Cavusoglu, G. M. Lee, and M. L. Z. Ma. It risk factor disclosure and stock price crashes. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[105] Statistics Solutions. Pearson correlation assumptions, 2023. Accessed: 24-March-2024.

[106] Jeremy C. Stein. Evaluating large-scale asset purchases, October 2012.

[107] S. Sunder. How did the u.s. stock market recover from the covid-19 contagion? *Mind Soc*, 20:261–263, 2021.

[108] K. Sheng Tai, R. Socher, and C.D. Manning. Improved semantic representations from tree-structured long short-term memory networks. http://arxiv.org/abs/1503.00075, 2015.

[109] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, 2015.

[110] P.C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168, 2007.

[111] The Economist. America's war on huawei nears its endgame. *The Economist*, July 2020.

[112] C. Ting-Fang and L. Li. Us-china tech war: Beijing's secret chipmaking champions. *Nikkei Asia*, May 2021.

[113] United States Securities and Exchange Commission. Form 10-k, 2024.

[114] G. Wang. The effects of quantitative easing announcements on the mortgage market: An event study approach. *International Journal of Financial Studies*, 7(1):9, 2019.

[115] Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, Ben Zhao, and Santa Barbara. Crowds on wall street: Extracting value from collaborative investing platforms. 03 2015.

[116] Z. Wang, Z. Hu, F. Li, et al. Learning-based stock trending prediction by incorporating technical indicators and social media sentiment. *Cognitive Computation*,

15:1092–1102, 2023. Received September 15, 2022; Accepted February 9, 2023; Published March 9, 2023; Issue Date May 1, 2023.

[117] R. Watts and J. Zimmerman. *Positive accounting theory*. Prentice Hall, Englewood Cliffs, NJ, 1986.

[118] Wiki. Form 10-k. https://en.wikipedia.org/wiki/Form_10-K, 2024. Wikipedia page.

[119] Wikipedia. Pearson correlation coefficient — Wikipedia, the free encyclopedia, 2024. [Online; accessed 24-March-2024].

[120] D. Wu, Y. Lee, and Y. Ngui. Chip shortage set to worsen as covid rampages through malaysia. *Bloomberg*, August 2021.

[121] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[122] H. You and Xj Zhang. Financial reporting complexity and investor underreaction to 10-k information. *Review of Accounting Studies*, 14:559–586, 2009.

[123] Yifan Yu. Us-china doomsday scenario not likely to happen, says nvidia's jensen huang. Nikkei Asian Review, 03 2024.

[124] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(e1253), 2018.

[125] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 649–657, 2015.

Appendix A

10-K Report Form

Part 1

Item 1: Business - This section describes the business of the company: its main products or services, subsidiaries, and markets in which it operates. It may also contain recent events, competition, regulation, and labour problems.

Item 1A: Risk Factors - This section provides risks and uncertainties, likely external effects, or possible failures that could affect their financial performance. Risk factors are generally enumerated based on their importance.

Item 1B: Unresolved Staff Comments - This section offers an explanation of any issues raised by the SEC staff on the previous reports if these issues have not been resolved afterwards.

Item 2:Properties - This section only lays out the company's significant physical properties, not intellectual or intangible property.

Item 3: Legal Proceedings - This section discloses any significant ongoing lawsuit or other legal proceeding.

Item 4 - [RESERVED]

Part 2

Item 5: Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities - This section discloses their performance in the stock market, dividends, and repurchases of their own stocks.

Item 6 - [RESERVED]

Item 7: Management's Discussion and Analysis of Financial Condition and Results of Operations (MDA) - This section discusses the firm's management for its financial performance, challenges, chances, and future outlook.

Item 7A: Quantitative and Qualitative Disclosures about Market Risks - This section shows the company's exposure to market risks.

Item 8: Financial Statement and Supplementary Data - This section offers the audited financial statements. It contains the balance sheet, income statement, cash flow statement, and footnotes.

Item 9: Changes in and Disagreements with Accountants on Accounting and Financial Disclosure - Companies address any alterations in or disputes with their accountants over financial reporting.

Item 9A: Controls and Procedures - This section details the company's procedures for disclosure controls and its internal control mechanisms for financial reporting.

Item 9B: Other Information - This section offers any additional information that does not align with the contents of other sections.

Part 3

Item 10: Directors, Executive Officers and Corporate Governance - This section delves into the specifics of the company's leadership, their respective roles, and the practices in place for corporate governance.

Item 11: Executive Compensation - This section addresses compensation policies and programmes, and the compensation of top executives.

Item 12: Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters - This section offers major shareholders' ownership of the company's stock as well as insiders.

Item 13: Certain Relationships and Related Transactions, and Director Independence- This section encompasses transactions involving directors, executives, and their affiliates, along with details regarding the independence of directors.

Item 14: Principal Accountant Fees and Services - The section details the fees charged by the company's auditors for their services.

Part 4

Item 15: Exhibits, Financial Statement Schedules - This section contains a list of the financial statements and exhibits.

[96, 118]

Appendix B

10-K Filing Extraction Model Example

B.1 10-K Filing (e.g Nvidia 2010-03-18)

...

[heading]Our Company[/heading]
NVIDIA Corporation helped awaken the world to the power of computer graphics when it invented the graphics processor unit, or GPU, in 1999. Expertise in programmable GPUs has led to breakthroughs in parallel processing which make supercomputing inexpensive and widely accessible.

...

[heading]ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS[/heading]
The following discussion and analysis of our financial condition and results of operations should be read in conjunction with "Item 1A. Risk Factors", "Item 6. Selected Financial Data", our Consolidated Financial Statements and related Notes thereto

...

[heading]Overview[/heading] [heading]Our Company[/heading]
NVIDIA Corporation helped awaken the world to the power of computer graphics when it invented the graphics processor unit, or GPU, in 1999. Expertise in programmable GPUs has led to breakthroughs in parallel processing which make supercomputing inexpensive and widely accessible. We serve the entertainment and consumer market with our GeForce graphics products, the professional design and visualization market with our Quadro graphics products, the high-performance computing market with our Tesla computing solutions products, and the mobile computing market with our Tegra system-on-a-chip products.

...

B.2 Risk Factor Section (e.g Nvidia 2023-02-24)

[title]Risk Factors Summary[/title]

[heading]Risks Related to Our Industry and Markets[/heading]

Failure to meet the evolving needs of our industry and markets may adversely impact our financial results. Competition in our current and target markets could cause us to lose market share and revenue.

[heading]Risks Related to Demand, Supply and Manufacturing[/heading]

• Failure to estimate customer demand properly has led and could lead to mismatches between supply and demand. • Dependency on third-party suppliers and their technology reduces our control over product quantity and quality, manufacturing yields, development, enhancement, and product delivery schedules and could harm our business. • Defects in our products have caused and could cause us to incur significant expenses to remediate and can damage our business.

...

Appendix C

Experimental Setup

C.1 Preprocessing

Through the 10-K filings extraction model at Chapter 3, we finally collected almost all 10-K filings listed in the QQQ from 2006 to 2023, which is 1383 filings as well as the text files with only the Item 1A risk factor section extracted. With these files, we preprocessed them before feeding them into our prediction model. Firstly, we split all sentences of a filing into words. Secondly, we removed non-alphabetic tokens from the texts such as numbers, proper nouns, and special characters(i.e. punctuations). For instance, Item 8 of a 10-K filing contains many numbers as a financial statement of a company is included in that section. As numbers were not necessary for textual analysis, we removed them. Thirdly, we removed stop words such as ”is”, ”the”, or ”and”, as they do not contain informative information. In the final step of the preprocessing, we selected lemmatisation, instead of stemming. Although lemmatisation is computationally more expensive than stemming, it tends to capture the more accurate base form of a word through linguistic analysis (i.e. considering Parts-of-speech) [7].

C.2 Hyper-parameter Setting

We set the hyper-parameters equal for the models to ensure a fair comparison between the sentiments with return labels and those with volatility labels for a sector level, a portfolio level, and a company level.

In Section 4.2, we replaced 0 by the θ in Equation 4.2 , and the value 0.5 by quantile q in Equation 4.3 . θ is a threshold and we used it to define high and low volatility. To figure out the balanced θ and q hyper-parameter for both the technology sector and a firm(in this paper, Nvidia), we selected the value of θ and q from Figure C.1, Figure C.2 on the training windows. In the technology sector from Figure C.1, we practically set θ as the 65th percentile of the distribution and q as 0.65. It implies we defined the volatility above the 65 quantile as high volatility, whereas the volatility below the 65 quantile is low volatility. Note that we set the 65th percentile and 0.65 for θ and q for both QQQ volatility itself and the volatilities of the top 10 firms in QQQ, respectively.

Empirically, the 3-day volatility for both QQQ and the top 10 showed a similar trend. We can see the peaks from the graph, referring to the financial crisis of 2008 and the COVID-19 pandemic around 2020. In the case of a firm(i.e. Nvidia) from Figure C.2, θ was set as the 65th percentile of the distribution, and q was 0.65. Similar to the QQQ volatility graph, Nvidia experienced high levels of volatility during the 2008 financial crisis and COVID-19 showed higher volatility movement in general during the training windows.

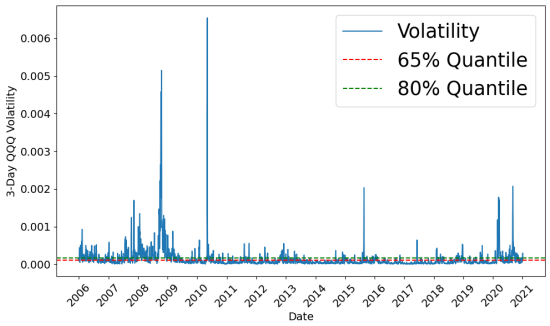

Figure C.1: 3-Day QQQ Volatility

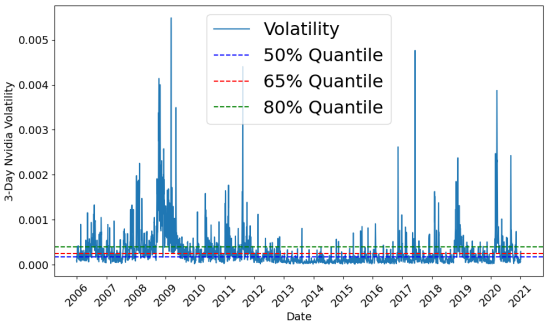

Figure C.2: 3-Day Nvidia Volatility

Furthermore, returns for both the technology sector and a firm showed a balanced proportion at the median of three-day returns. Hence, we set 0.5 at Equation 4.3 for both the technology sector in Figure C.3 and a firm return in Figure C.4.

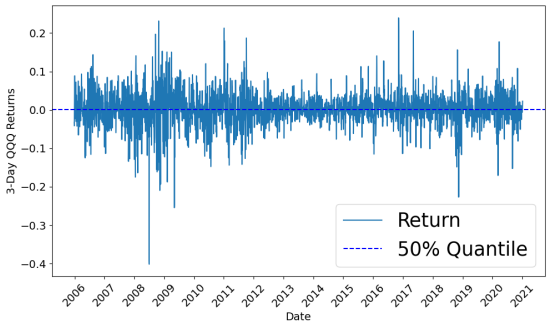

Figure C.3: 3-Day QQQ Return

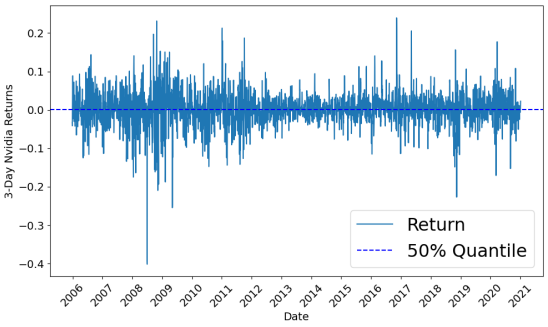

Figure C.4: 3-Day Nvidia Return

In Equation 4.3, we specifically defined α and k. Note that α was a threshold to filter out sentiment-neutral words. We set α^+ and α^- within the interval $(0, 0.5]$ such that each of the positive word sets and the negative word sets includes 100 words. Moreover, note that $k \in N$ was another threshold to relate to the count of word w across all filings such that we used k as a minimum frequency requirement to reduce the influence of rare words. We set k as the 90 percentile quantile of the term frequency distribution. It means we ignored words that appear less than the 90 percentile quantile in a filing.

In Equation 4.8, we also defined λ. It was a positive constant and used in a penalty term to adjust our model. The penalty term was used to avoid the model overfitting when few sentiment-charged words appear in the filing. Without the penalty term, the models can consider the filing that contains few negative words but does not include positive as a negative filing. However, just because the model contains few negative

sentiment-charged words without positive words, that does not mean the filing has a negative tone. To control this phenomenon, we set the penalty coefficient λ to 0.1 in the penalty term.

C.3 Baseline

The purpose of the paper was to predict the sentiment score for both a technology sector and a firm from the contents of 10-K filings. To achieve that, we could infer the sentiment scores for $\hat{p}^{RET}$ and $\hat{p}^{VOL}$ by labelling with return and volatility, respectively. Then, we introduced a baseline sentiment score to compare our sentiment scores with it.

Our baseline model to calculate baseline sentiment scores was suggested by [39], and used the dictionary created by [73]. Our baseline Loughran and McDonald(LM) sentiment score,$\hat{p}^{LM}$, is computed as

$$\hat{p}_i^{LM} = \left(\sum_{w \in LM^+} d_{i,w} - \sum_{w \in LM^-} d_{i,w} \right) \left(\sum_{w \in V} d_{i,w} \right)^{-1}, \tag{C.1}$$

where LM^+ refers to the positive word lists and LM^- refers to the negative words list of Loughran and McDonald's dictionary. To attain the base sentiment scores, we calculated the difference between the number of positive words and the number of negative words, and then we divided this result by the total number of words in each 10-K filing.

To explain more of the LM's dictionary for our robust evaluation, the LM dictionary is traditionally used for financial analysis. LM offered an improved textual dictionary for a better accurate financial analysis in financial documents. Existing the financial dictionary of 10-K filings based on the Harvard dictionary misclassified the negative words in a financial context. Three-fourths of negative words in the 10-K filings do not carry a negative connotation in financial reports. To solve this issue, LM suggested three approaches. Firstly, LM created a refined list of words that more accurately reflects negative sentiment in a financial context, by analysing every word that appears in at least 5% of the SEC's 10-K filings. Secondly, LM introduced a term weighting scheme that controls the influence of frequently mentioned words and amplifies the significance of rarer terms, thus mitigating the misclassification of words. Finally, they added five other word classifications(e.g., positive, uncertainty, litigious, strong modal, and weak modal words). They found that these new classifications can be linked to market reactions, volatility in stock returns, unexpected earnings, and trading volumes [73]. In our study, we used only negative and positive categories to adjust the financial dictionary for our model.

C.4 Portfolio

In this study, we formed a portfolio to evaluate the technology sector sentiment scores. A portfolio can be changed to any portfolio for financial analysis. In practice of our study, we selected Invesco QQQ Trust Series 1 an exchange-traded fund (QQQ or

QQQ ETF). This passive fund(i.e., our portfolio) tracks the Nasdaq 100 Index, which consists of shares from 100 of the largest and most innovative non-financial firms listed on the Nasdaq stock exchange. Holdings in QQQ are predominantly in large-cap technology firms, accounting for 60% of the portfolio. As such, the QQQ is conventionally considered as a technology sector fund. The top 10 holdings represent a 50% allocation of the portfolio, with 9 out of 10 firms being in the tech sector. To represent the technology sector in the US, we formed two portfolios from the QQQ fund. Firstly, we formed the portfolio, which has the exact same allocation proportion as the QQQ fund itself as of 2023. This allocation proportion is annually corrected so that our current portfolio reflects 2023 allocation data. The actual 2023 portfolio allocation can be found in Appendix D. This portfolio is used to compare the sector sentiment scores.

The second portfolio was constructed with the top 10 firms, considering the asset allocation ratio of the first portfolio. In other words, the second portfolio consisted of the top 10 firms, accounting for 50% of the QQQ portfolio, according to the proportion invested in the first portfolio. This portfolio was computed as:

$$A_j = \frac{w_j}{\sum_{j=1}^{10} w_j}, \tag{C.2}$$

where j denoted a firm invested at the j-th ranking in the 2023 portfolio Appendix D, and w_j represented the portfolio weight of the j-th firm. A_j referred to the allocated proportion of the j-th firm in the 2023 portfolio.

In the case of a single firm evaluation, we did not form a portfolio for it. Instead, we followed the firm's stock market price.

Appendix D

Invesco QQQ Trust, Series 1 Portfolio [5]

Ticker	CIK	Weight	Ticker	CIK	Weight	Ticker	CIK	Weight
AAPL	320193	9.217	ADP	8670	0.758	AEP	4904	0.337
MSFT	789019	8.548	PANW	1327567	0.752	AZN	901832	0.307
AMZN	1018724	4.867	REGN	872589	0.722	TEAM	1650372	0.301
AVGO	1730168	4.192	KLAC	319201	0.637	BIIB	875045	0.3
META	1326801	3.846	MELI	1099590	0.632	CTSH	1058290	0.3
TSLA	1318605	3.79	SNPS	883241	0.627	CEG	1868275	0.298
NVDA	1045810	3.736	CDNS	813672	0.593	FAST	815556	0.297
GOOGL	1652044	2.571	CSX	277948	0.548	DDOG	1561550	0.296
GOOG	1652044	2.51	PYPL	1633917	0.535	DASH	1792789	0.294
COST	909832	2.316	ASML	937966	0.531	MRNA	1682852	0.294
ADBE	796343	2.16	MAR	1048286	0.52	EA	712515	0.293
PEP	77476	1.841	CTAS	723254	0.487	ON	1097864	0.292
AMD	2488	1.837	LULU	1397187	0.487	CSGP	1057352	0.283
NFLX	1065280	1.705	ABNB	1774585	0.478	GEHC	1932393	0.282
INTC	50863	1.688	NXPI	1413447	0.474	EXC	1109357	0.28
CSCO	858877	1.62	MNST	865752	0.468	BKR	1701605	0.277
TMUS	1283699	1.438	CRWD	1535527	0.465	VRSK	1442145	0.272
CMCSA	1166691	1.399	ROP	882835	0.46	XEL	72903	0.27
INTU	896878	1.387	CHTR	1091667	0.457	GFS	1709048	0.269
QCOM	804328	1.284	WDAY	1327811	0.454	ZS	1713683	0.264
TXN	97476	1.23	ORLY	898173	0.442	TTD	1671933	0.26
AMGN	318154	1.205	MRVL	1835632	0.418	ANSS	1013462	0.249
AMAT	6951	1.09	ADSK	769397	0.415	CDW	1402057	0.243
HON	773840	1.087	PCAR	75362	0.405	DLTR	935703	0.242
BKNG	1075531	0.987	MCHP	827054	0.392	CCEP	1650107	0.24
ISRG	1035267	0.942	DXCM	1093557	0.378	MDB	1441816	0.236
SBUX	829224	0.862	CPRT	900075	0.373	FANG	1539838	0.226
VRTX	875320	0.833	ROST	745732	0.368	WBD	1437107	0.222
LRCX	707549	0.831	IDXX	874716	0.366	TTWO	946581	0.218
GILD	882095	0.79	KDP	1418135	0.366	SPLK	1353283	0.203
ADI	6281	0.783	FTNT	1262039	0.364	WBA	1618921	0.182
MDLZ	1103982	0.774	ODFL	878927	0.36	ILMN	1110803	0.177
PDD	1737806	0.769	KHC	1637459	0.355	SIRI	908937	0.168
MU	723125	0.762	PAYX	723531	0.344			

Appendix E

Pearson Correlation

E.1 Definition

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{E.1}$$

- r refers to the correlation coefficient.
- x_i and y_i refer to the individual sample points for variables x and y, respectively.
- $\bar{x}$ and $\bar{y}$ represent the mean values of the samples for x and y.

[119]

Appendix F

Sentiment Score Prediction Model Results

F.1 The Sector Sentiment Model with Only-Risk-Factor (Figure F.1)

F.2 The Top10 Sentiment Model with 10-K (Figure F.2)

F.3 The Top10 Sentiment Model with Only-Risk-Factor (Figure F.3)

F.4 Nvidia Sentiment Model with 10-K (Figure F.4)

F.5 Nvidia Sentiment Model with Only-Risk-Factor (Figure F.5)

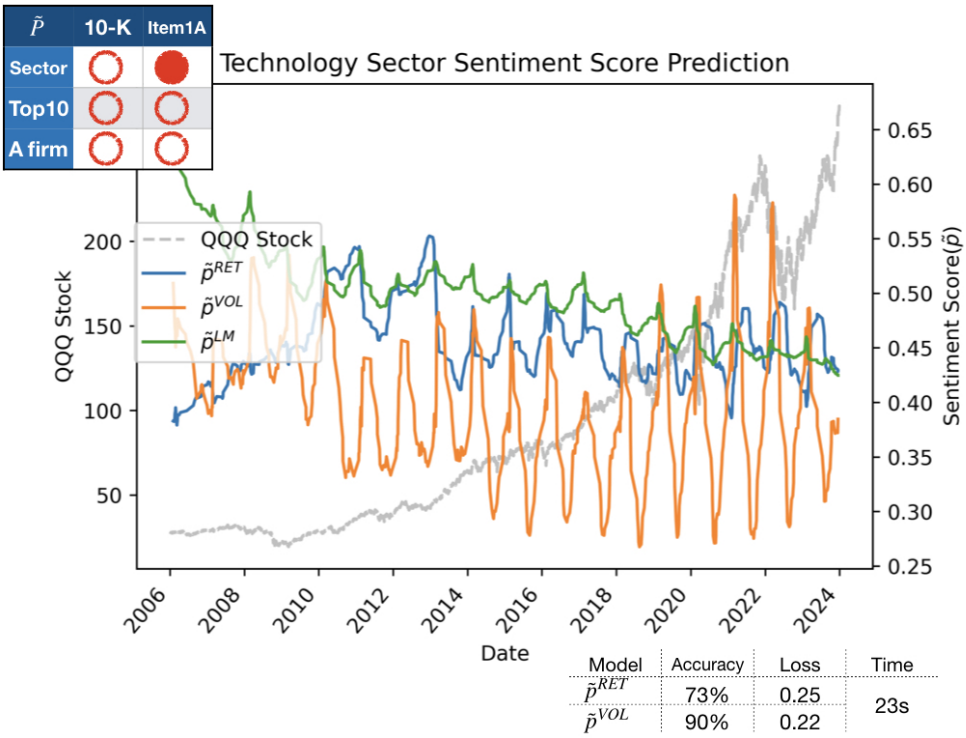

Figure F.1: The Sector Sentiment Model with Only-Risk-Factor

$r_{\tilde{p}_i,\tilde{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	**-0.167	1		
LM	0.008	**0.182	1	
Stock	**-0.376	*0.056	**-0.764	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table F.1: Filtered, A Sector Sentiment Correlation with Only-Risk-Factor

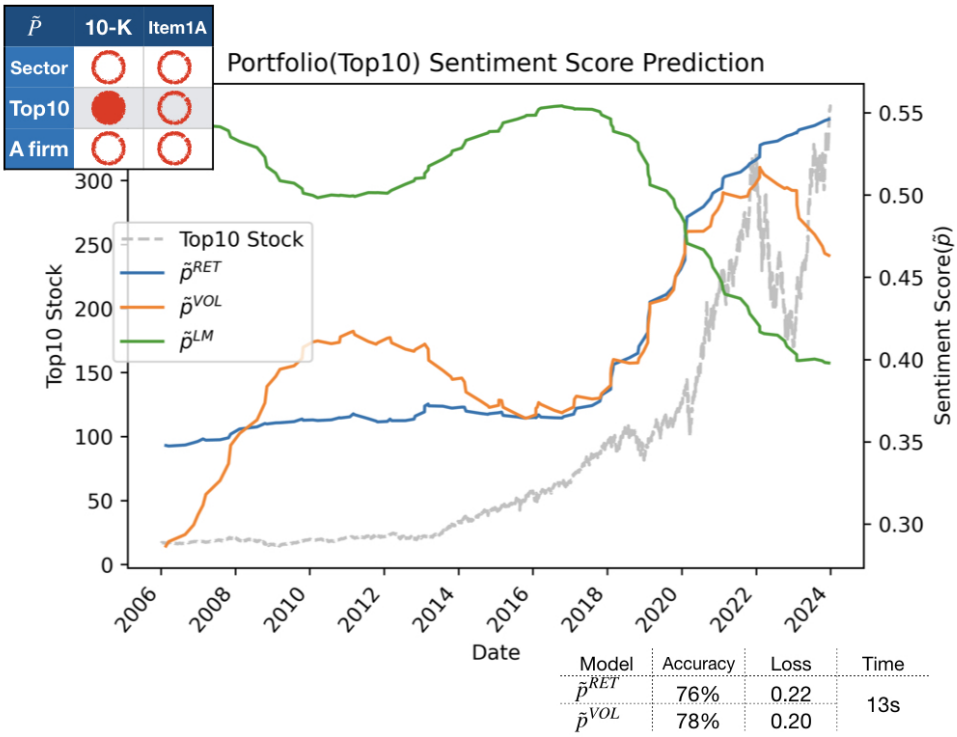

Figure F.2: The Top10 Sentiment Model with 10-K filing

$r_{\tilde{p}_i,\tilde{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	**0.905	1		
LM	**-0.922	**-0.903	1	
Stock	**0.956	**0.824	**-0.841	1

Note :* $p\text{-}value < 0.05$,** $p\text{-}value < 0.005$

Table F.2: Filtered, The Top10 Sentiment Correlation with 10-K filing

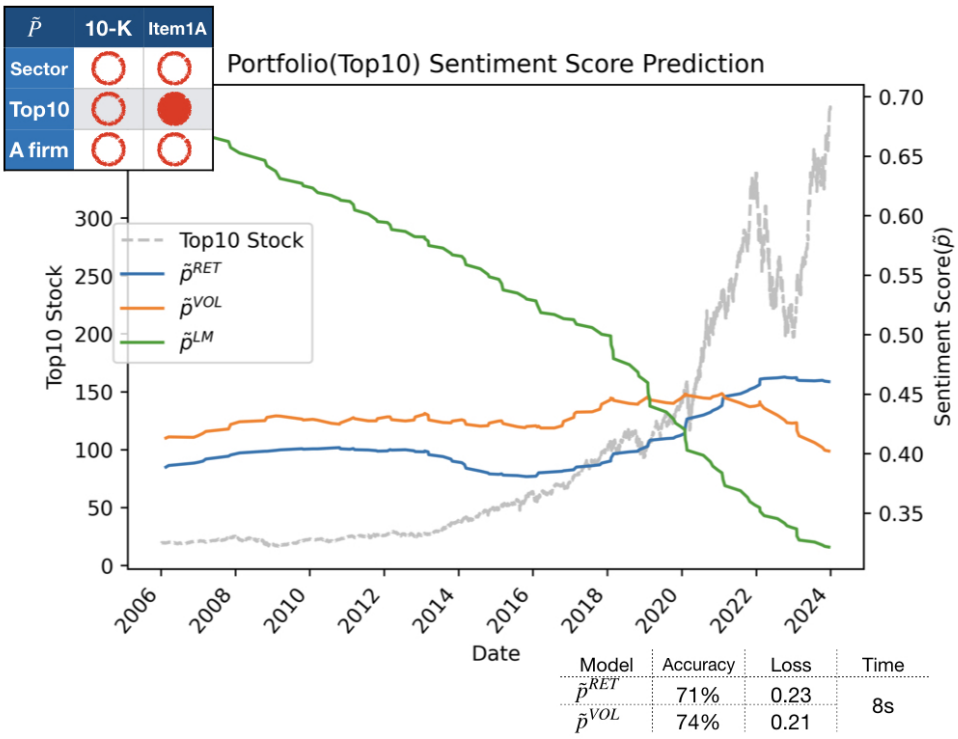

Figure F.3: The Top10 Sentiment Model with Only-Risk-Factor

$r_{\tilde{p}_i,\tilde{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	0.145	1		
LM	**-0.806	**-0.349	1	
Stock	**0.893	*0.185	**-0.937	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table F.3: Filtered, The Top10 Sentiment Correlation with Only-Risk-Factor

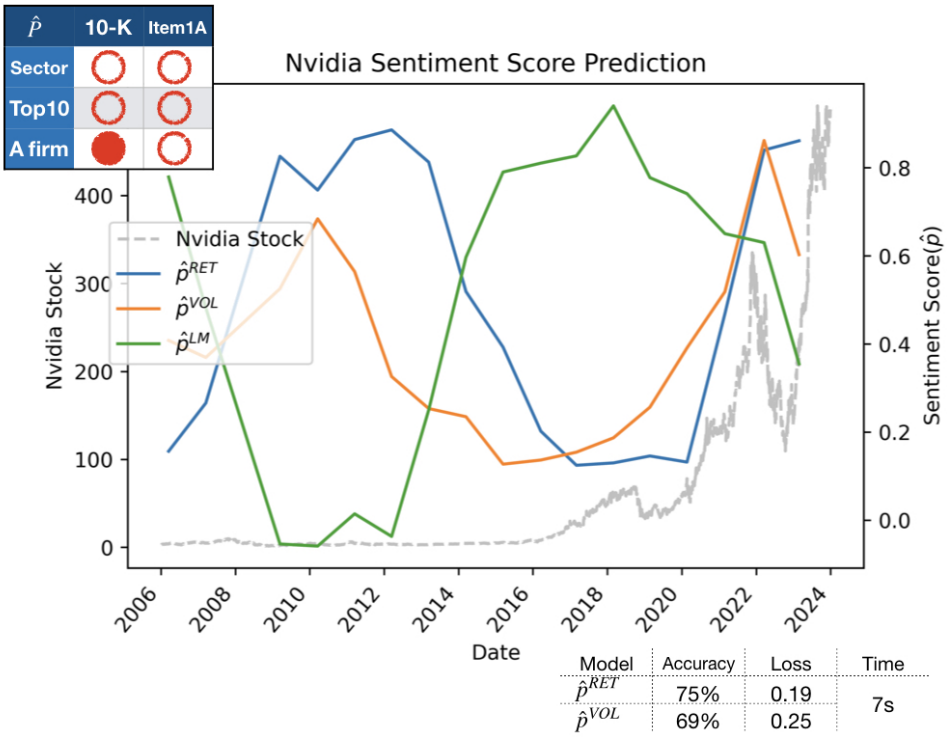

Figure F.4: Nvidia Sentiment Model with 10-K filing

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	*0.612	1		
LM	**-0.832	*-0.482	1	
Stock	0.234	0.588	0.182	1

Note : * $p\text{-value} < 0.05$, ** $p\text{-value} < 0.005$

Table F.4: Nvidia Sentiment Correlation with 10-K filing

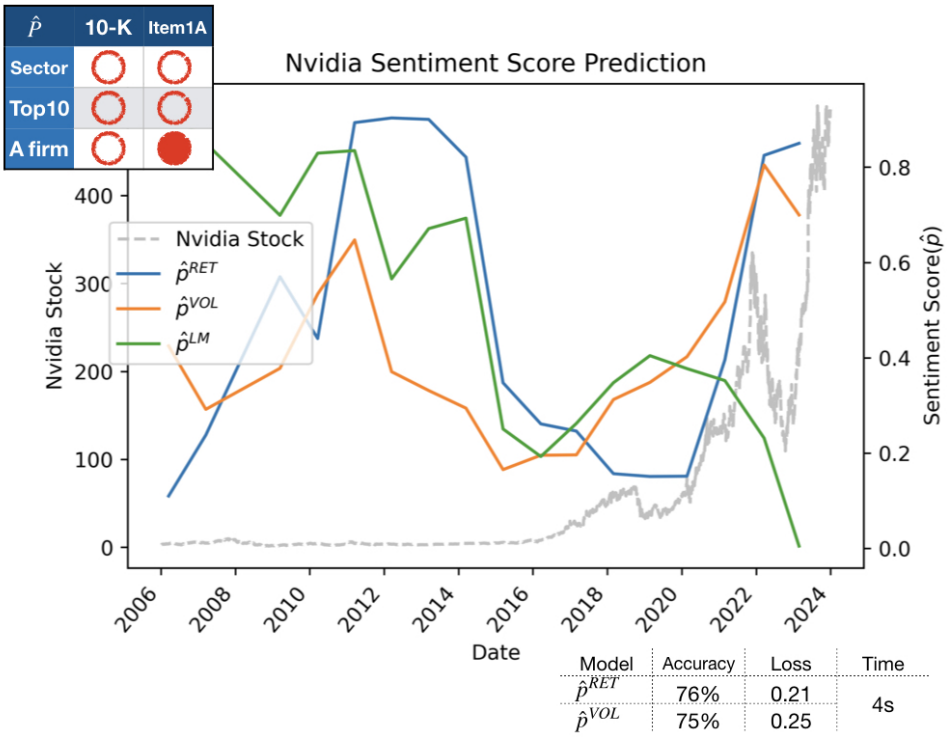

Figure F.5: Nvidia Sentiment Model with Only-Risk-Factor

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	$^{*}0.490$	1		
LM	0.057	0.004	1	
Stock	0.228	$^{**}0.712$	*-0.593	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table F.5: Nvidia Sentiment Correlation with Only-Risk-Factor

Appendix G

Unfiltered Sentiment Prediction Scores

G.0.1 The Unfiltered Sector Sentiment Model with 10-K

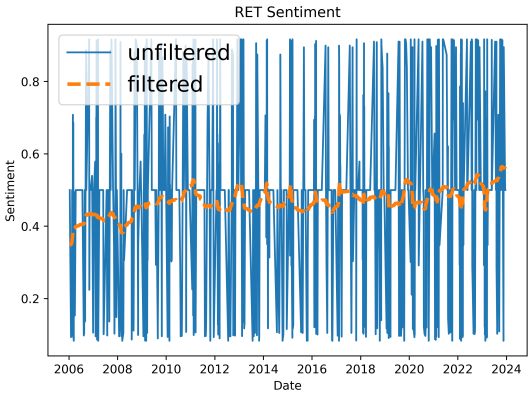

Figure G.1: Unfiltered Sector p^{RET} with 10-K

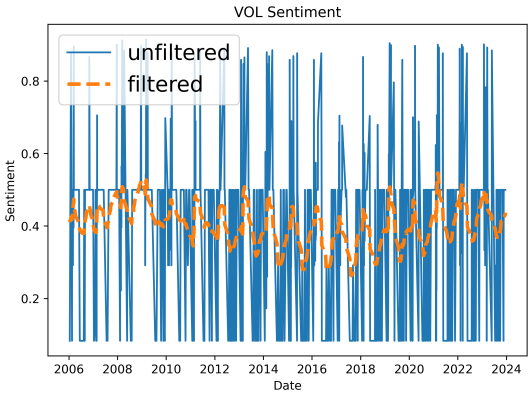

Figure G.2: Unfiltered Sector p^{VOL} with 10-K

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	*0.045	1		
LM	*-0.049	**-0.225	1	
Stock	*-0.001	*0.041	**-0.270	1

Note :* $p\text{-}value < 0.05$,** $p\text{-}value < 0.005$

Table G.1: Unfiltered, A Sector Sentiment Correlation with 10-K filing

G.0.2 The Unfiltered Sector Sentiment Model with Only-Risk-Factor

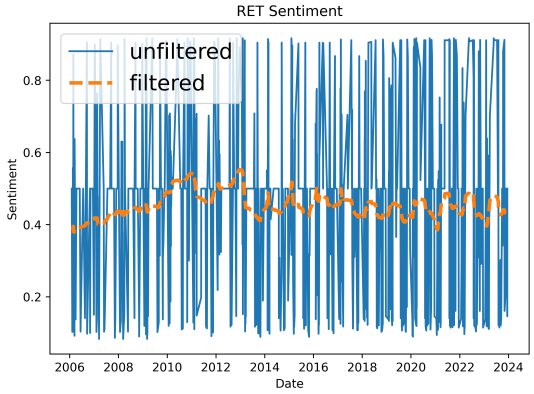

Figure G.3: Unfiltered Sector p^{RET}

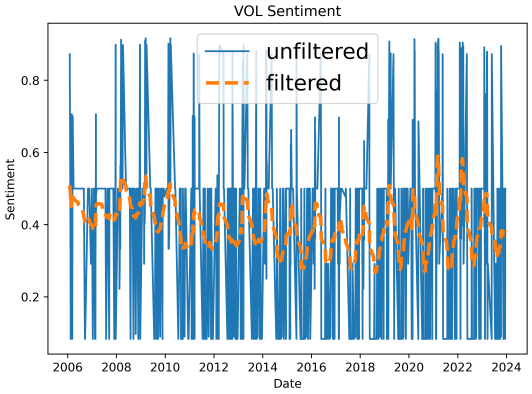

Figure G.4: Unfiltered Sector p^{VOL}

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	**-0.081	1		
LM	0.033	**0.103	1	
Stock	*-0.071	0.030	**-0.248	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table G.2: Unfiltered, The Sector Sentiment Correlation with Only-Risk-Factor

G.0.3 The Unfiltered Top10 Sentiment Model with 10-K filing

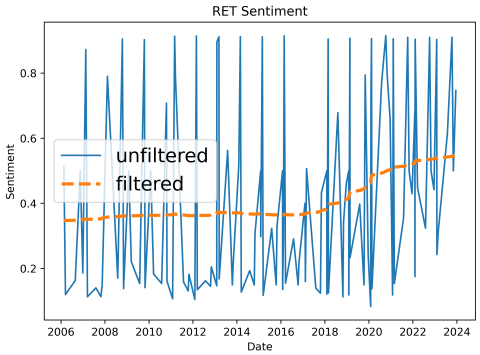

Figure G.5: Unfiltered Top10 p^{RET}

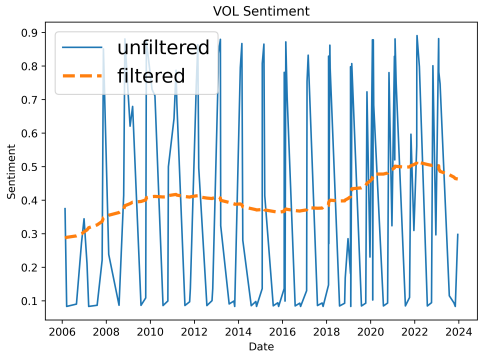

Figure G.6: Unfiltered Top10 p^{VOL}

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	0.153	1		
LM	0.829	**-0.349	1	
Stock	*0.267	0.117	*-0.220	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table G.3: Unfiltered, The Top10 Sentiment Model with 10-K filing

G.0.4 The Unfiltered Top10 Sentiment Model with Only-Risk-Factor

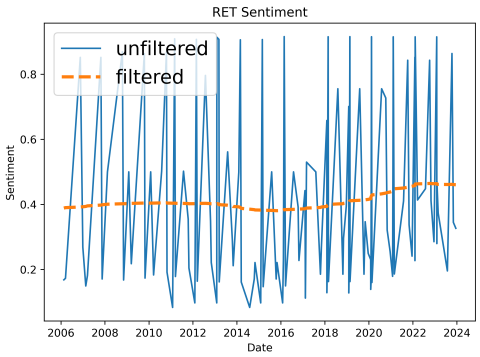

Figure G.7: Unfiltered Top10 p^{RET}

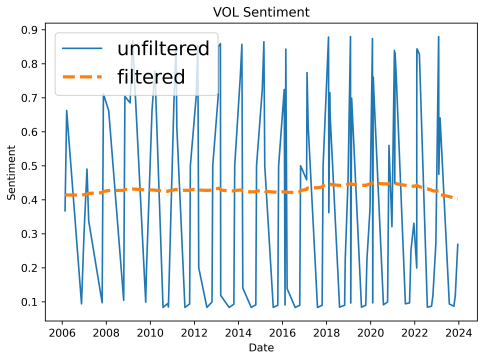

Figure G.8: Unfiltered Top10 p^{VOL}

$r_{\hat{p}_i,\hat{p}_j}$	RET	VOL	LM	Stock
RET	1			
VOL	0.037	1		
LM	0.202	$^{**}0.414$	1	
Stock	0.105	-0.021	**-0.447	1

$Note: ^{*}\ p\text{-}value < 0.05, ^{**}\ p\text{-}value < 0.005$

Table G.4: Unfiltered, The Top10 Sentiment Correlation with Only-Risk-Factor

Appendix H

Data Pipeline Automation

Our system can update the latest SEC filings, facilitating data preparation(i.e. preprocessing, extracting the risk factor section, and creating a document-term dictionary). Basically, our system can collect every type of SEC filing. In practice, we collected 10-K filings, followed by the extraction of the risk factor section and the creation of a document-term dictionary. As mentioned in Experiment Result, we generated sentiment metrics for three different stakeholders with either the entire 10-K filing or the risk factor section, respectively. To do that, we need 6 types of a document-term dictionary. One of the dictionaries, for instance, an individual firm(e.g. Nvidia)'s document-term dictionary from the complete 10-K filings.

Our system can automatically create 6 types of a document-term dictionary with the latest 10-K filings. This automation process works on Apache Airflow, an open-source platform designed to manage workflow processes in data engineering pipelines. Our airflow automation system consists of three dags corresponding to each stakeholder level. Note that a Directed Acyclic Graph(DAG) refers to a collection of all the takes you want to execute, arranged to reflect their relationships and dependencies. Each dag creates the corresponding document-term dictionary for our Sentiment Score Prediction Model. As mentioned in Project Objective, the publication dates of 10-K filing are various to each firm although it should be released nearly every day throughout that year. Thus, we scheduled our dags differently. Sector-level dag(i.e. Technology sector from the QQQ) updates daily, Portfolio-level dags(i.e. Top10) updates monthly, and Firm-level dag update yearly. You can check data workflow in Figure H.1

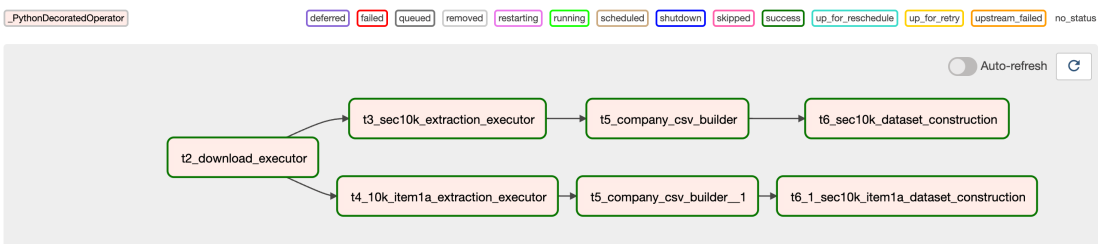

Figure H.1: Airflow Dags