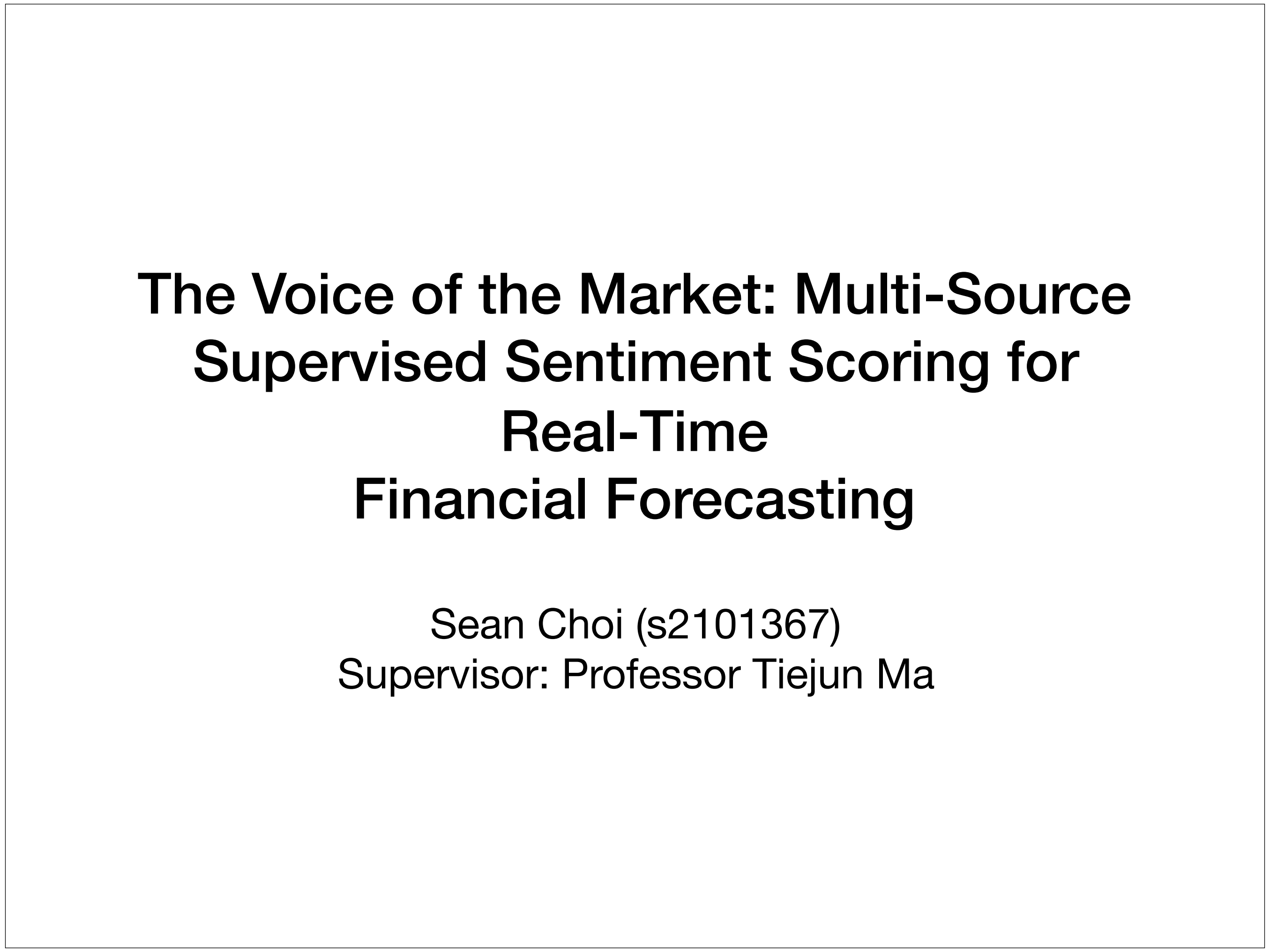
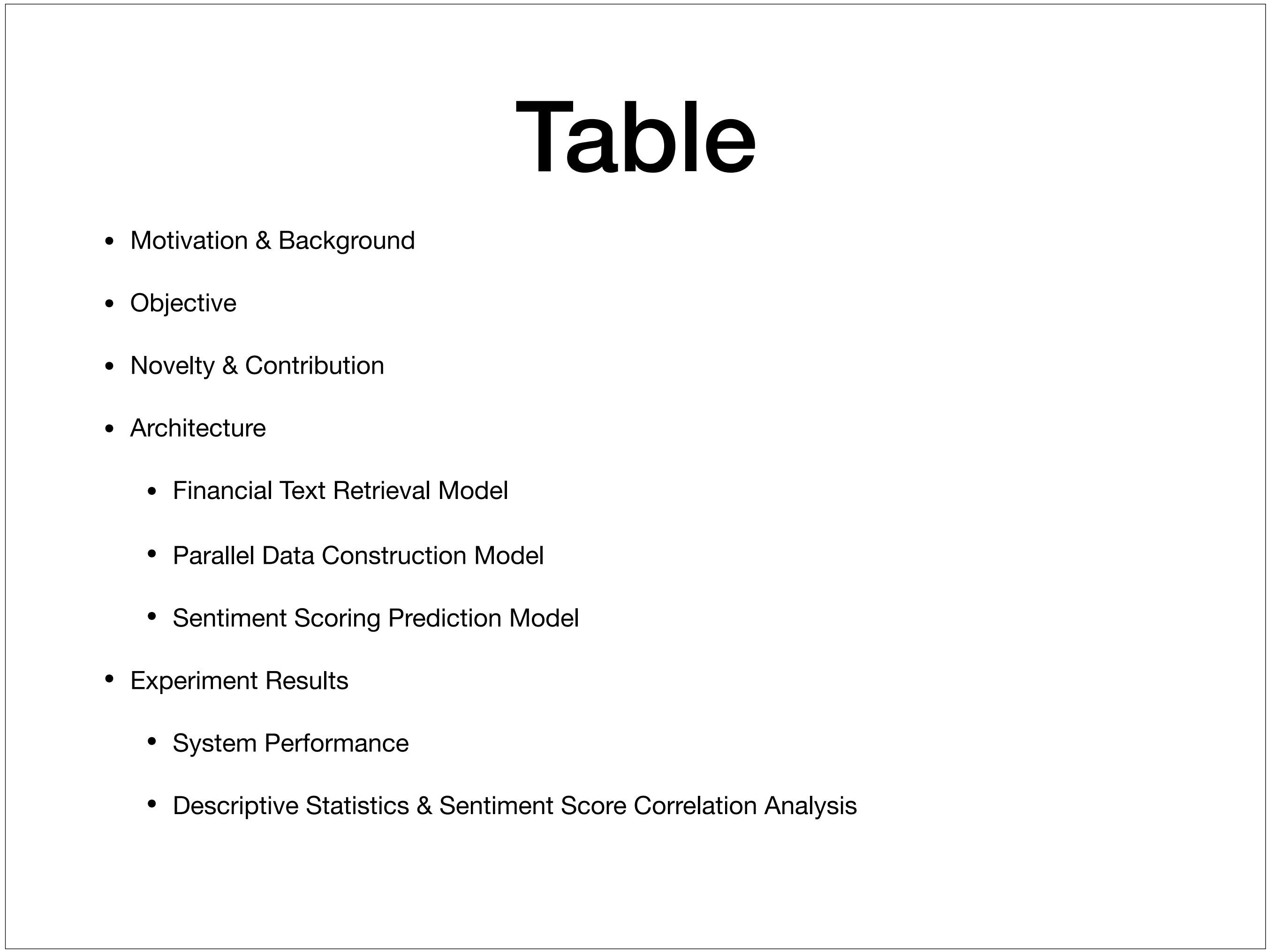


The Voice of the Market: Multi-Source Supervised Sentiment Scoring for Real-Time Financial Forecasting

Sean Choi (s2101367)
Supervisor: Professor Tiejun Ma

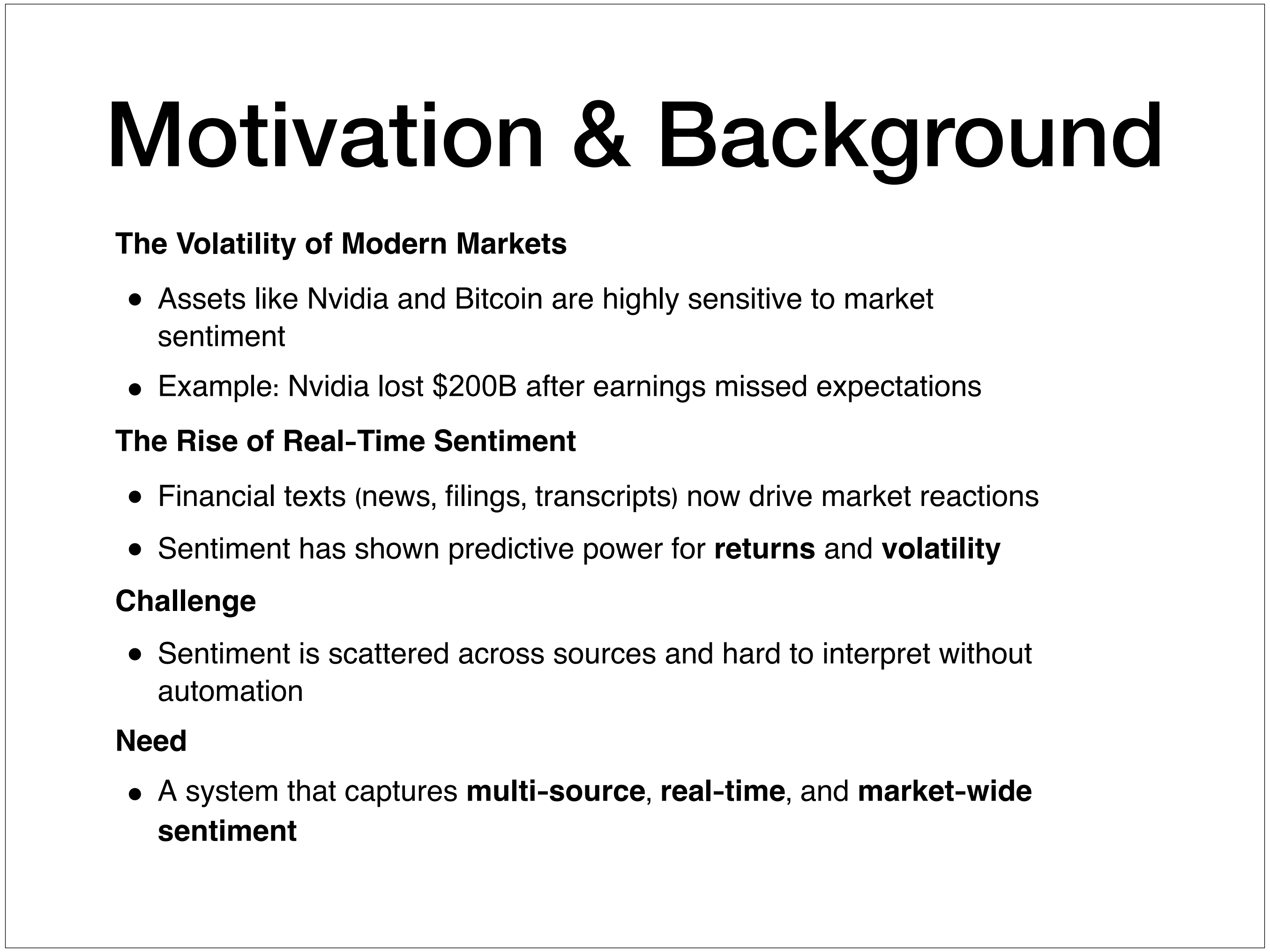
Good [morning/afternoon], everyone. My name is Sanggyu Sean Choi, and today, I'll be presenting my Master's thesis titled:
"The Voice of the Market: Multi-Source Supervised Sentiment Scoring for Real-Time Financial Forecasting."



Table

- Motivation & Background
- Objective
- Novelty & Contribution
- Architecture
  - Financial Text Retrieval Model
  - Parallel Data Construction Model
  - Sentiment Scoring Prediction Model
- Experiment Results
  - System Performance
  - Descriptive Statistics & Sentiment Score Correlation Analysis

My 10 to 15 min presentation will consist of background explanation, goal, key contribution, Architecture, and Experiment Result.



Motivation & Background

The Volatility of Modern Markets

- Assets like Nvidia and Bitcoin are highly sensitive to market sentiment
- Example: Nvidia lost $200B after earnings missed expectations

The Rise of Real-Time Sentiment

- Financial texts (news, filings, transcripts) now drive market reactions
- Sentiment has shown predictive power for **returns** and **volatility**

Challenge

- Sentiment is scattered across sources and hard to interpret without automation

Need

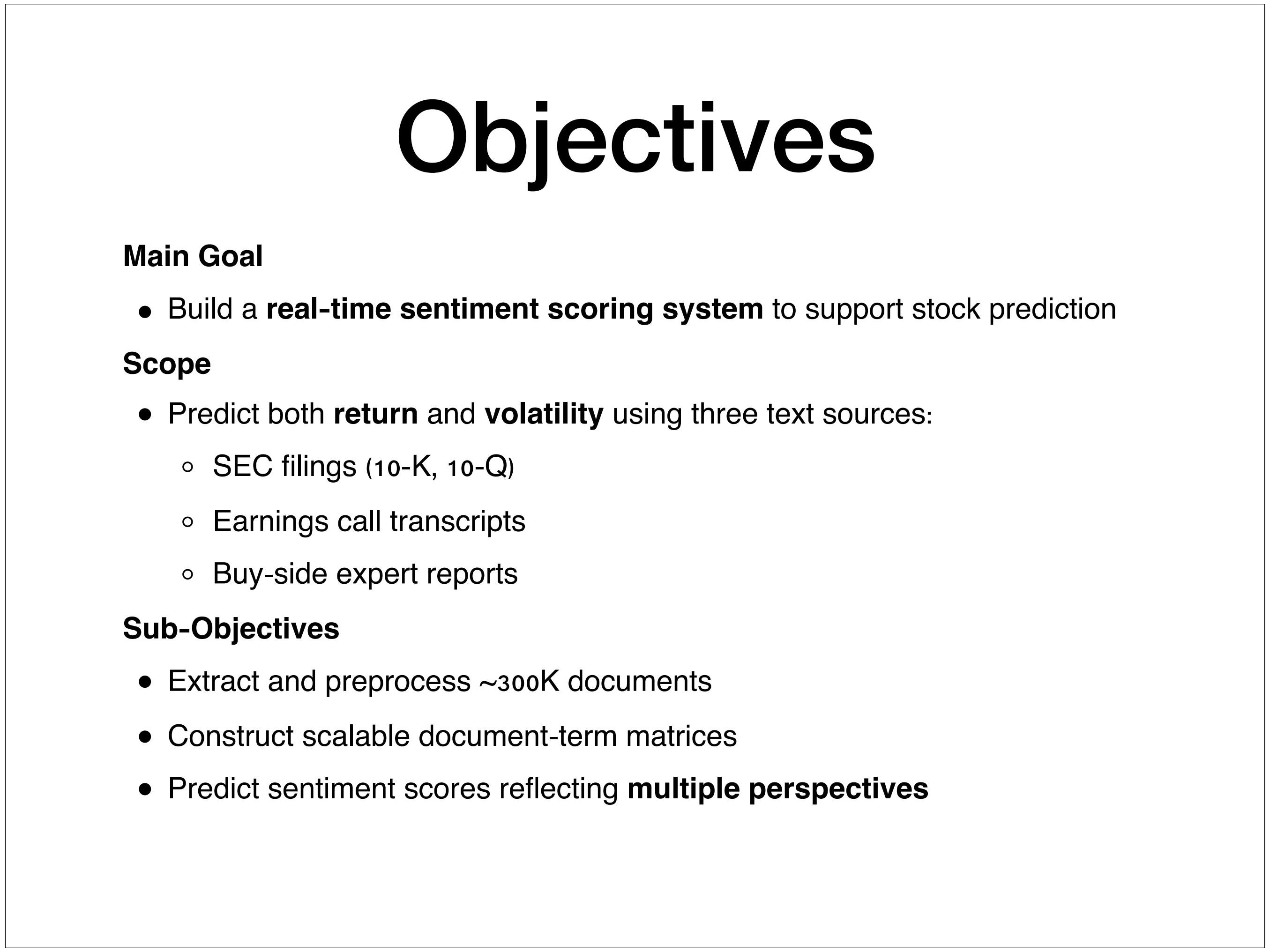
- A system that captures **multi-source**, **real-time**, and **market-wide sentiment**

Let me start with the motivation behind this work.

As a retail investor with a five-year track record and a Compound Annual Growth Rate of 7.5%, I’ve experienced firsthand the difficulty of navigating volatile assets like Nvidia and Bitcoin. For example, Nvidia lost nearly $200 billion in market value after a single earnings report. This highlights a key insight: market movements today are highly driven by real-time sentiment—not just fundamentals.

In modern finance, sentiment travels faster than fundamentals. From news, social media, and earnings calls, we now have a multi-layered, instant ecosystem of information that significantly affects returns and volatility. Academic research and industry practice both confirm that real-time sentiment signals can lead to better, more informed trading decisions.

Yet, there exists no comprehensive, real-time system that integrates multiple sources of financial text into actionable sentiment metrics. That’s the gap this project addresses.



Objectives

Main Goal

- Build a **real-time sentiment scoring system** to support stock prediction

Scope

- Predict both **return** and **volatility** using three text sources:
    - SEC filings (10-K, 10-Q)
    - Earnings call transcripts
    - Buy-side expert reports

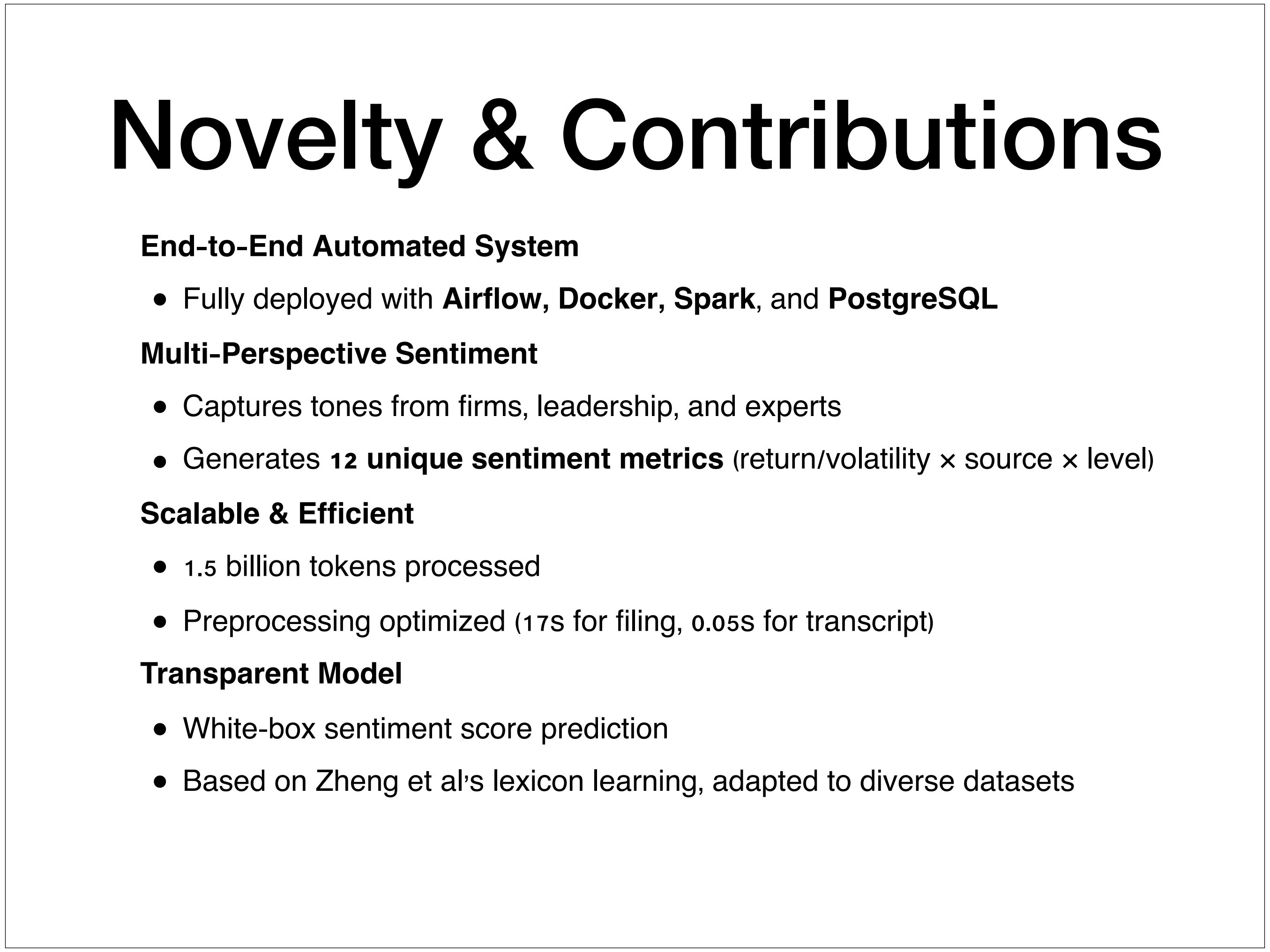
Sub-Objectives

- Extract and preprocess ~300K documents
- Construct scalable document-term matrices
- Predict sentiment scores reflecting **multiple perspectives**

The main objective of my thesis is to build a real-time, scalable system that automatically predicts sentiment scores for the US stock market using three major types of financial texts:

SEC Filings (10-K and 10-Q)
Earning Call Transcripts
Buy-Side Analyst Reports

Each represents a different voice in the market—the official firm stance, leadership tone, and expert analysis. These sentiment scores are designed to predict both returns and volatility across firm-level and market-level views.



Novelty & Contributions

End-to-End Automated System

- Fully deployed with **Airflow, Docker, Spark**, and **PostgreSQL**

Multi-Perspective Sentiment

- Captures tones from firms, leadership, and experts
- Generates **12 unique sentiment metrics** (return/volatility × source × level)

Scalable & Efficient

- 1.5 billion tokens processed
- Preprocessing optimized (17s for filing, 0.05s for transcript)

Transparent Model

- White-box sentiment score prediction
- Based on Zheng et al’s lexicon learning, adapted to diverse datasets

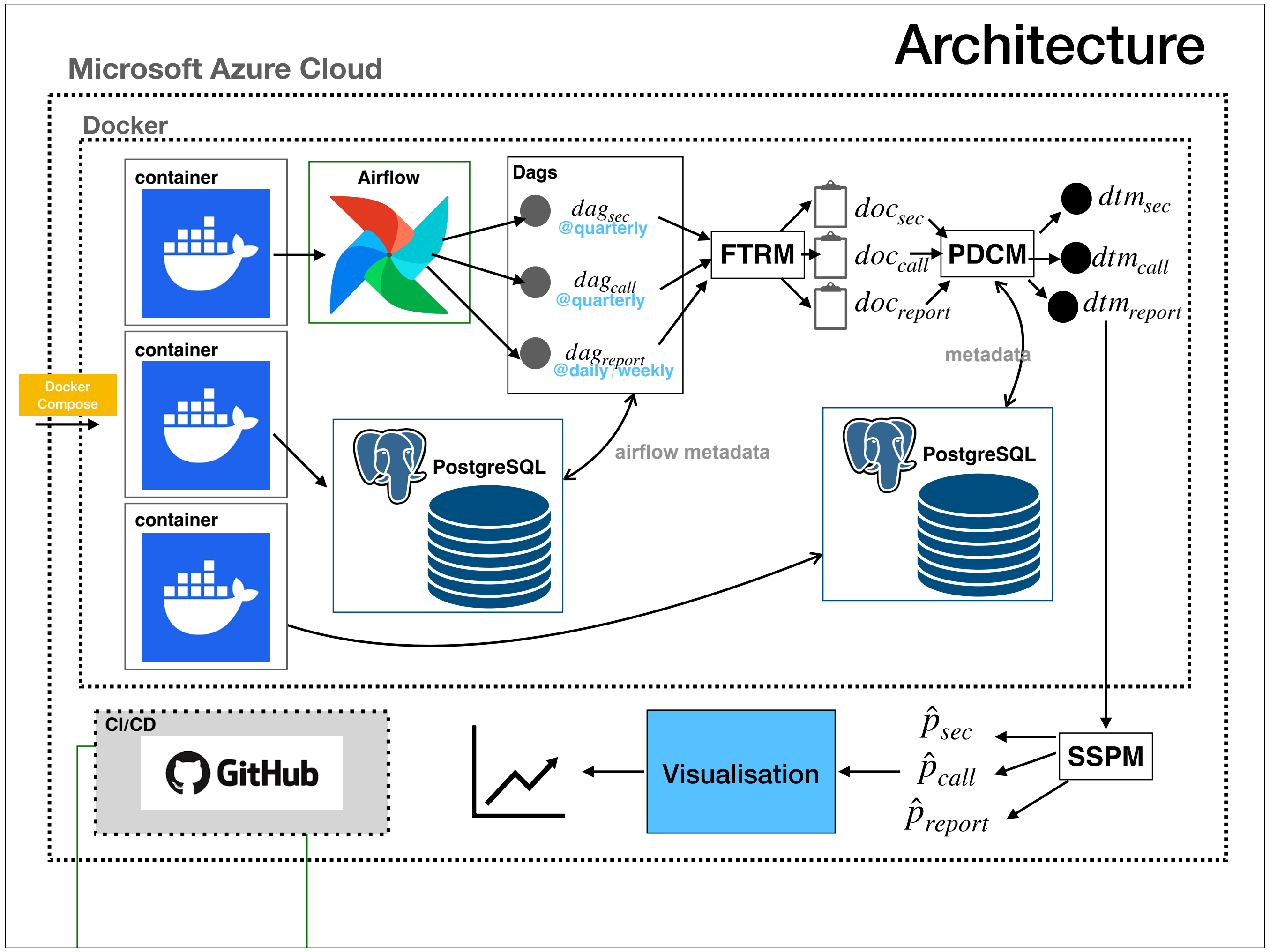
This work contributes to the field in several novel ways:

First, I developed an end-to-end real-time system, deployed using Docker and orchestrated by Apache Airflow, which handles data retrieval, processing, and sentiment prediction.

Second, the sentiment score prediction model is based on a scalable and transparent econometric approach, as proposed by the reference paper, but applied to diverse datasets—not just news—and extended to generate both return and volatility predictive signals.

Third, no existing system processes SEC filings, earnings calls, and analyst reports in real-time for market-level and company-level sentiment scores. This is the first to do so at scale.

Architecture



In this research, I built an end-to-end automated system for generating real-time sentiment metrics. The architecture integrates three major components:

Financial Text Retrieval Model (FTRM) – Automatically gathers texts from EDGAR, Seeking Alpha, and the Ninja API. Over 300,000 documents were retrieved covering S&P 500 firms from 2006 to 2025.
Parallel Data Construction Model (PDCM) – Preprocesses documents and builds document-term metrics using Apache Spark and Parquet. It efficiently handles over 1.5 billion tokens using distributed processing.
Sentiment Score Prediction Model (SSPM) – Predicts sentiment scores for each document based on co-occurrence and supervised lexicon learning, producing both return-predictive and volatility-predictive signals.

These components are **orchestrated under Apache Airflow**, ensuring each step—from data retrieval to sentiment prediction—runs seamlessly and automatically as soon as new financial text is available.

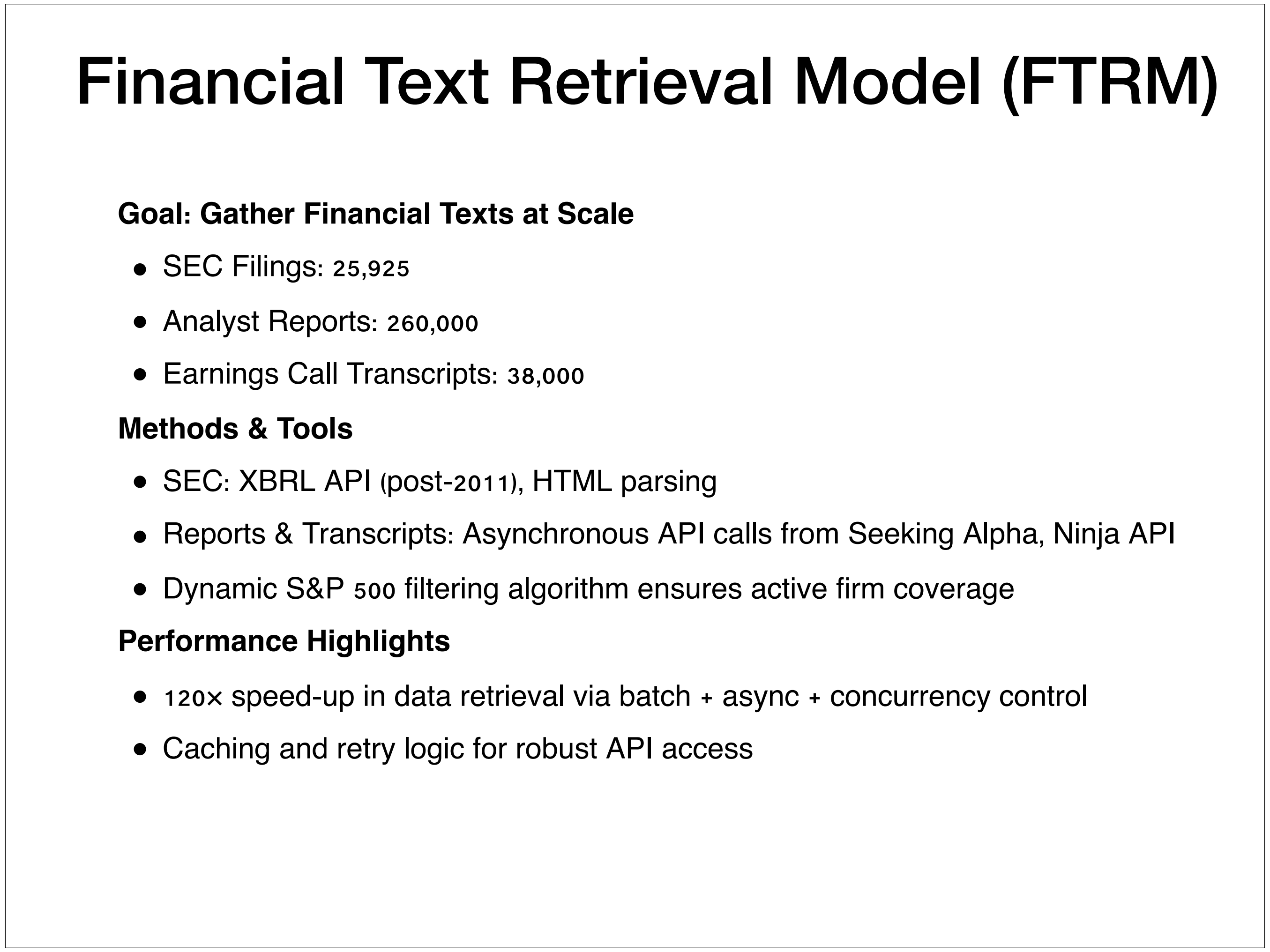
The system is deployed on **Microsoft Azure Cloud**, which ensures scalability, and it's **containerised using Docker,** making it modular, portable, and easy to upgrade.

We handle three types of financial text:
SEC filings (10-K and 10-Q),
Earnings call transcripts, and
Buy-side expert reports.

Each document type triggers a dedicated **Airflow DAG,** which processes the document and generates both **market-level** and **firm-level** sentiment scores. And for each, we compute scores for return and volatility—yielding a total of **12 sentiment metrics.**

Custom DAG scheduling accounts for the **irregular nature** of financial publications—since not all reports are released on the same schedule.

To ensure responsiveness, the pipeline integrates with **PostgreSQL** for metadata tracking and uses **Continuous Integration and Continuous Delivery via GitHub** for continuous improvement and reliability.



Financial Text Retrieval Model (FTRM)

Goal: Gather Financial Texts at Scale

- SEC Filings: 25,925
- Analyst Reports: 260,000
- Earnings Call Transcripts: 38,000

Methods & Tools

- SEC: XBRL API (post-2011), HTML parsing
- Reports & Transcripts: Asynchronous API calls from Seeking Alpha, Ninja API
- Dynamic S&P 500 filtering algorithm ensures active firm coverage

Performance Highlights

- 120× speed-up in data retrieval via batch + async + concurrency control
- Caching and retry logic for robust API access

Let me introduce the first module of the system: the Financial Text Retrieval Model (FTRM).

This model is responsible for collecting financial texts across three diverse sources to reflect three distinct market perspectives:

SEC Filings (10-K and 10-Q):

Represent the firm’s official tone and regulatory stance.
We retrieved over 25,925 filings using a modernised XBRL API-based algorithm.
This replaced the outdated RSS feed method, allowing access to filings in structured JSON format and reducing retrieval latency.

Buy-Side Analyst Reports:

Capture the voice of financial experts and portfolio managers.
Collected from Seeking Alpha, a leading crowdsourced platform, yielding 260,000 reports.
A novel parallel asynchronous API strategy with multiple threads concurrency made the retrieval 120× faster, slashing collection time from 8000 to ~67 hours.

Earnings Call Transcripts:

Reflect management tone and leadership sentiment.
Retrieved 38,000+ transcripts using the same high-speed method.

To ensure data validity, we dynamically tracked active firms in the S&P 500 from 2006 to 2025, using metadata like permanent identifiers and name end dates from Wharton Research Data Services (WRDS).
This allowed us to handle mergers, ticker changes, and rebrandings accurately—something many academic systems overlook.

The FTRM is highly robust, generalizable, and production-ready for expanding to new text types like social media or regulatory press releases.

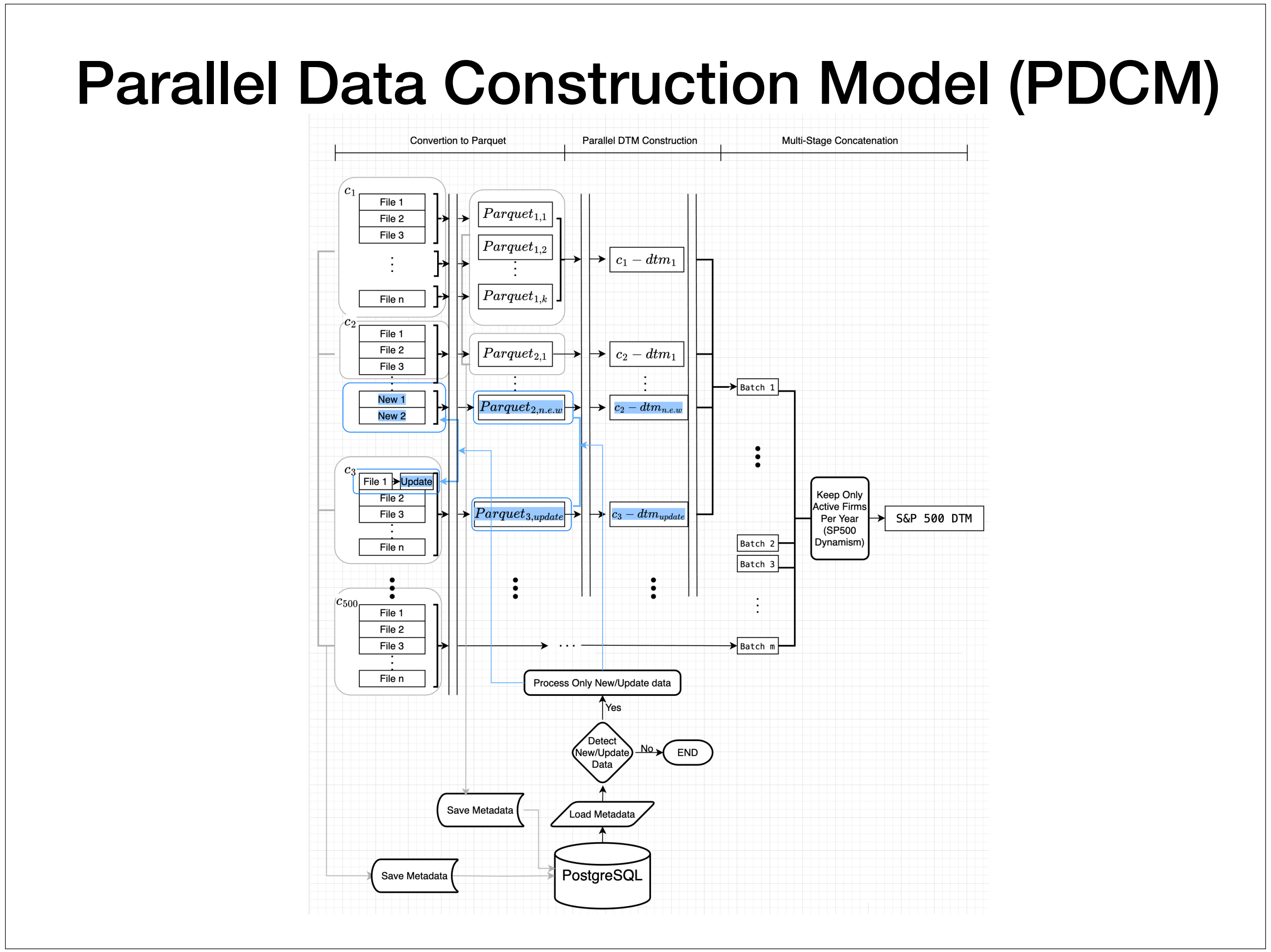
The FTRM fetches three types of documents:

25,925 SEC Filings: Retrieved using the new XBRL API.

260,000 Analyst Reports: Collected from Seeking Alpha using asynchronous API calls.

38,000 Earning Call Transcripts: Retrieved and parsed using the same efficient algorithm.

Each document is timestamped, filtered, and formatted for downstream processing.



Parallel Data Construction Model (PDCM)

Once the data is collected, it flows into the Parallel Data Construction Model (PDCM)—our second core module.

This model handles two critical tasks:
Efficient preprocessing of documents, and
Construction of document-term metrics (DTMs) for sentiment modeling.

We’re dealing with ~300,000 medium-to-large documents and 1.3 billion tokens. So, real-time preprocessing must be extremely fast and scalable. Here's how we handled that:

Step 1: Conversion to Apache Parquet
Raw HTML and JSON files are cleaned and transformed into columnar Parquet format.
This allows efficient I/O operations, reduces memory usage, and integrates smoothly with Spark DataFrames for distributed processing.

Step 2: Parallel DTM Construction
Using Apache Spark’s RDD architecture, we construct one DTM per firm in parallel.
These DTMs are stored in Parquet and only updated when files are newly added or modified, thanks to metadata tracking with PostgreSQL.

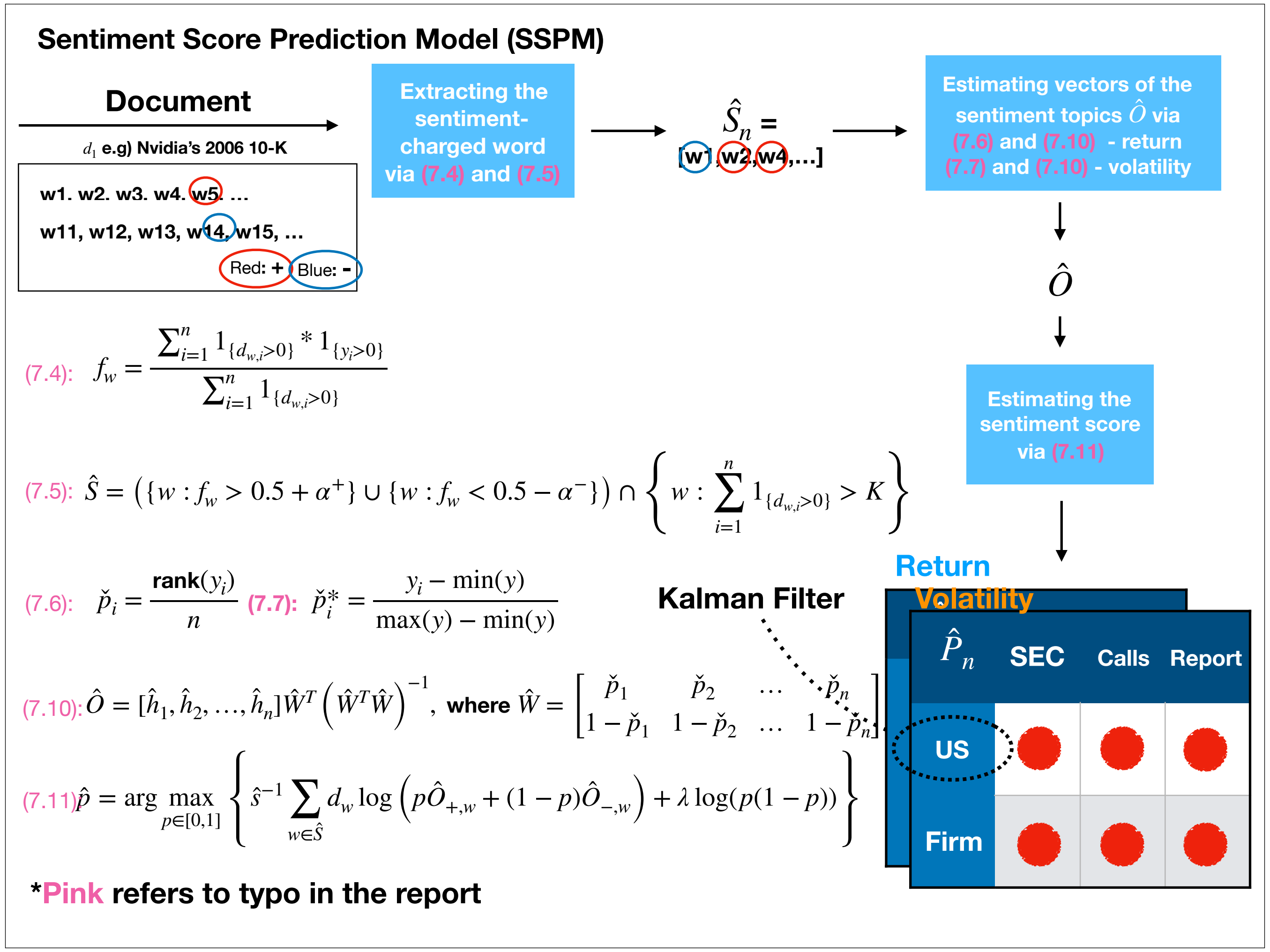
Step 3: Multi-Stage Concatenation
Once firm-level DTMs are built, we aggregate them into market-level matrices using a scalable multi-stage concatenation logic.
This makes sentiment computation at the US market scale not just feasible—but efficient and real-time.

This parallel construction reduced preprocessing latency by a factor of 5×, turning a traditionally batch-heavy task into a near real-time engine.

Given the scale of the data, traditional preprocessing methods were insufficient. The PDCM includes:

- Conversion to Parquet for columnar storage and faster I/O
- Spark-based Parallel Document-Term Matrix Construction
- Metadata-aware updates, avoiding reprocessing unchanged files

This reduced training time by 5x and made real-time predictions feasible.



Sentiment Score Prediction Model (SSPM)

Document
d_1 e.g) Nvidia's 2006 10-K
w1. w2. w3. w4. w5. …
w11, w12, w13, w14, w15, …
Red: +
Blue: -
Extracting the sentiment-charged word via (7.4) and (7.5)
$\hat{S}_n$ =
[w1,w2,w4,…]
Estimating vectors of the sentiment topics $\hat{O}$ via (7.6) and (7.10) - return (7.7) and (7.10) - volatility
$\hat{O}$
Estimating the sentiment score via (7.11)
Kalman Filter
Return
Volatility

$\hat{P}_n$	SEC	Calls	Report
US	●	●	●
Firm	●	●	●

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

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

(7.6): $$\check{p}_i = \frac{\textbf{rank}(y_i)}{n}$$ **(7.7):** $$\check{p}_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}$$

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

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

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

The Sentiment Score Prediction Model(SSPM) uses automatic labeling—mapping stock returns and volatility to textual patterns—to learn a sentiment lexicon. It's fully transparent and avoids the black-box problem of deep learning.

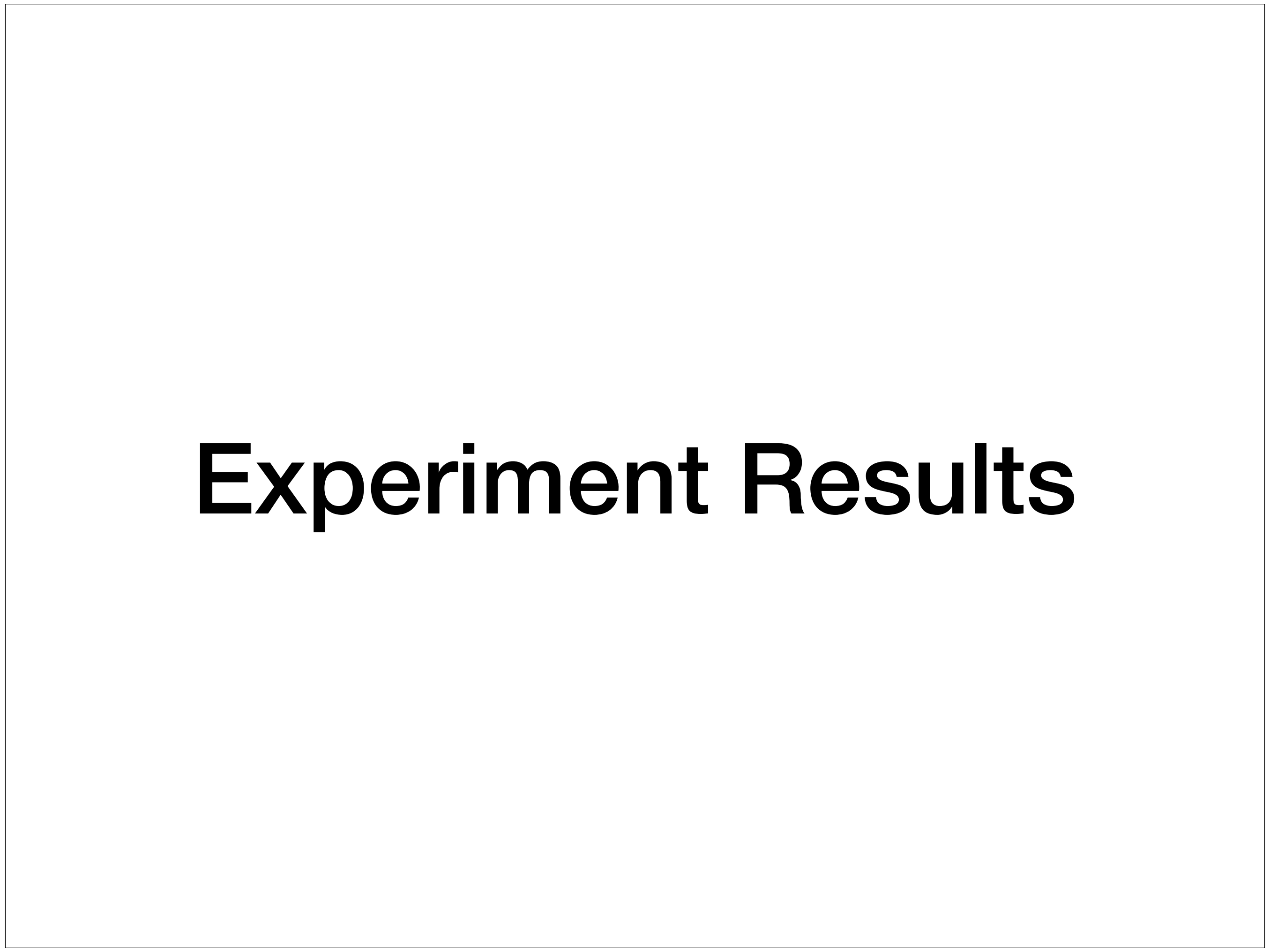
Sentiment scores are generated at market and company levels, reflecting:
Firm's own statements
Leadership's tone in calls
Financial expert opinions

In total, 12 sentiment metrics were generated
I will briefly explain it in today's presentation. You can find more details in our report.

After extracting a variety of types of documents from them, we will filter the sentiment-charged words via equations (7.4), and (7.5). Equation (7.4) defines the frequency of each word w in documents with positive returns compared to its overall frequency in all reports. Equation (7.5) is about identifying sentiment-charged words. It involves setting thresholds to determine whether a word is positively or negatively charged in sentiment. We filter out the set of neutral words through this equation. If f_w equals 0.5, it means the word w has a neutral tone, but we remove the set of natural words to remove the noise.

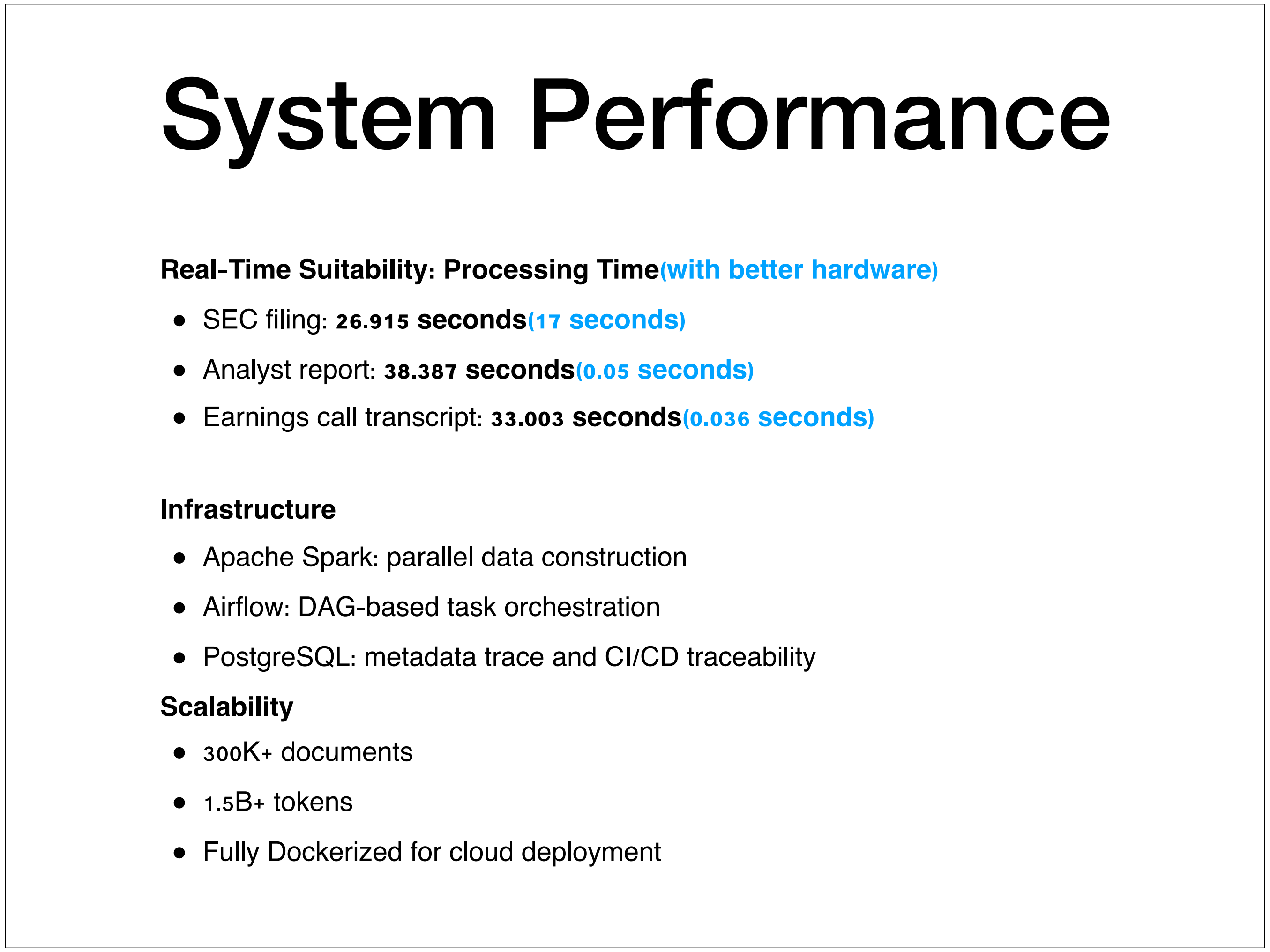
After attaining the sentiment-charged-words for a single report, we will estimate vectors of the sentiment topics via equations (7.6) and (7.10) or (7.7) and (7.10). To estimate the sentiment topic vector, both return and volatility, respectively, we should use different sentiment proxies (that is, the p inverted circumflex). The 7.6 sentiment proxy is for return, and the 7.7 sentiment proxy is for volatility. The pink refers to a typo in Figure 7.1 and Chapter 7 in the report. This slide version of the figure is correct. In this chapter 7 of the report, the related equations and technologies were mistakenly labelled as 6.x, but they should be 7.x. Please keep this in mind while reading the report.

Please check this slide when you read the report, if you need more clarification.



Experiment Results

The following part is experiment results of our study



System Performance

Real-Time Suitability: Processing Time(with better hardware)

- SEC filing: **26.915 seconds**(17 seconds)
- Analyst report: **38.387 seconds**(0.05 seconds)
- Earnings call transcript: **33.003 seconds**(0.036 seconds)

Infrastructure

- Apache Spark: parallel data construction
- Airflow: DAG-based task orchestration
- PostgreSQL: metadata trace and CI/CD traceability

Scalability

- 300K+ documents
- 1.5B+ tokens
- Fully Dockerized for cloud deployment

Next, let’s look at the **performance** of the system in a real-world scenario.

One of the major contributions of this research is demonstrating that **real-time sentiment generation** is practical—even for massive financial text datasets.

Here’s how fast the pipeline runs with only four virtualised cores 2.45 GHz CPU with 15 RAM:
SEC filing: 26.915 seconds
Analyst report: 38.387 seconds
Earnings call transcript: 33.003 seconds

This speed is achieved through several architectural optimizations:
Distributed Spark Processing in both preprocessing and scoring
Dockerized Microservices ensure modular, lightweight, and isolated tasks
PostgreSQL Metadata Tracking allows intelligent incremental updates
Airflow DAG Orchestration handles time-sensitive triggers for each document type

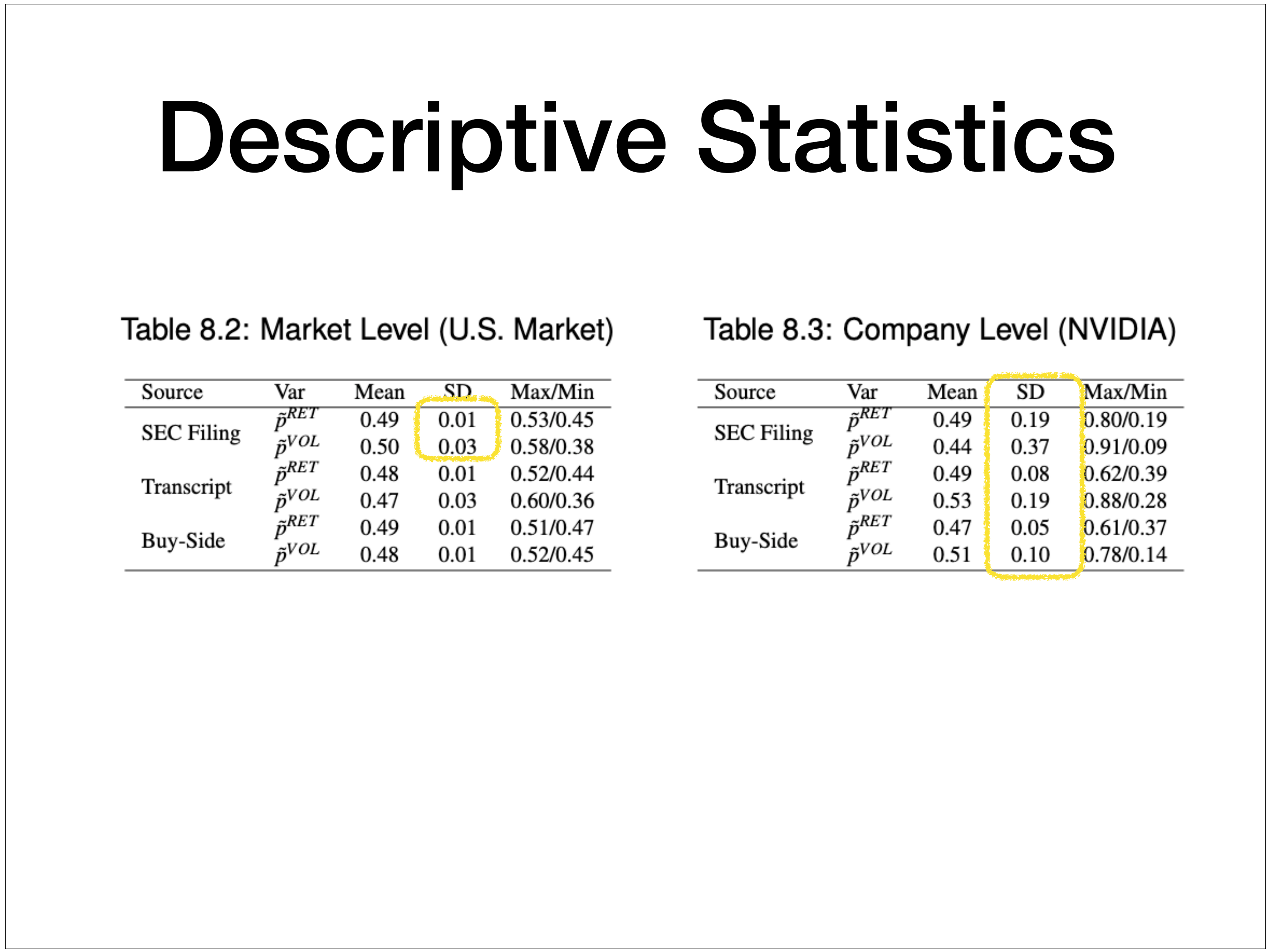
Even with **1.5 billion tokens,** our system can compute real-time sentiment scores within seconds of document release—making it suitable for high-frequency strategies or live dashboards. Additionally, The average 30 seconds of performance for real-time case might not be strong performance as the performance is affected by the hardware specification, and system performance improved significantly with better hardware. In the different test with better hardware of a 6-core 3.3 GHz CPU and 24 GB RAM, processing times dropped to 17s for SEC filings, 50 milliseconds for transcripts, and just 0.36 milliseconds for articles. The containerised, cloud-native design ensures the system can scale effortlessly with future hardware upgrades.

In summary, the system is scalable, robust, and ready for production in real-world financial environments.

Performance is a key metric.

SEC filing processed: 17 seconds
Analyst report: 0.036 seconds
Earnings call: 0.05 seconds

This allows the system to keep up with high-frequency document releases, crucial for real-time strategies.
System performance improved significantly with better hardware. In the different test with better hardware of a 6-core 3.3 GHz CPU and 16 GB RAM, processing times dropped to 17s for SEC filings, 0.05s for transcripts, and just 0.0036s for articles. The containerised, cloud-native design ensures the system can scale effortlessly with future hardware upgrades.



Descriptive Statistics

Table 8.2: Market Level (U.S. Market)

Source	Var	Mean	SD	Max/Min
SEC Filing	$\bar{p}^{RET}$	0.49	0.01	0.53/0.45
	$\bar{p}^{VOL}$	0.50	0.03	0.58/0.38
Transcript	$\bar{p}^{RET}$	0.48	0.01	0.52/0.44
	$\bar{p}^{VOL}$	0.47	0.03	0.60/0.36
Buy-Side	$\bar{p}^{RET}$	0.49	0.01	0.51/0.47
	$\bar{p}^{VOL}$	0.48	0.01	0.52/0.45

Table 8.3: Company Level (NVIDIA)

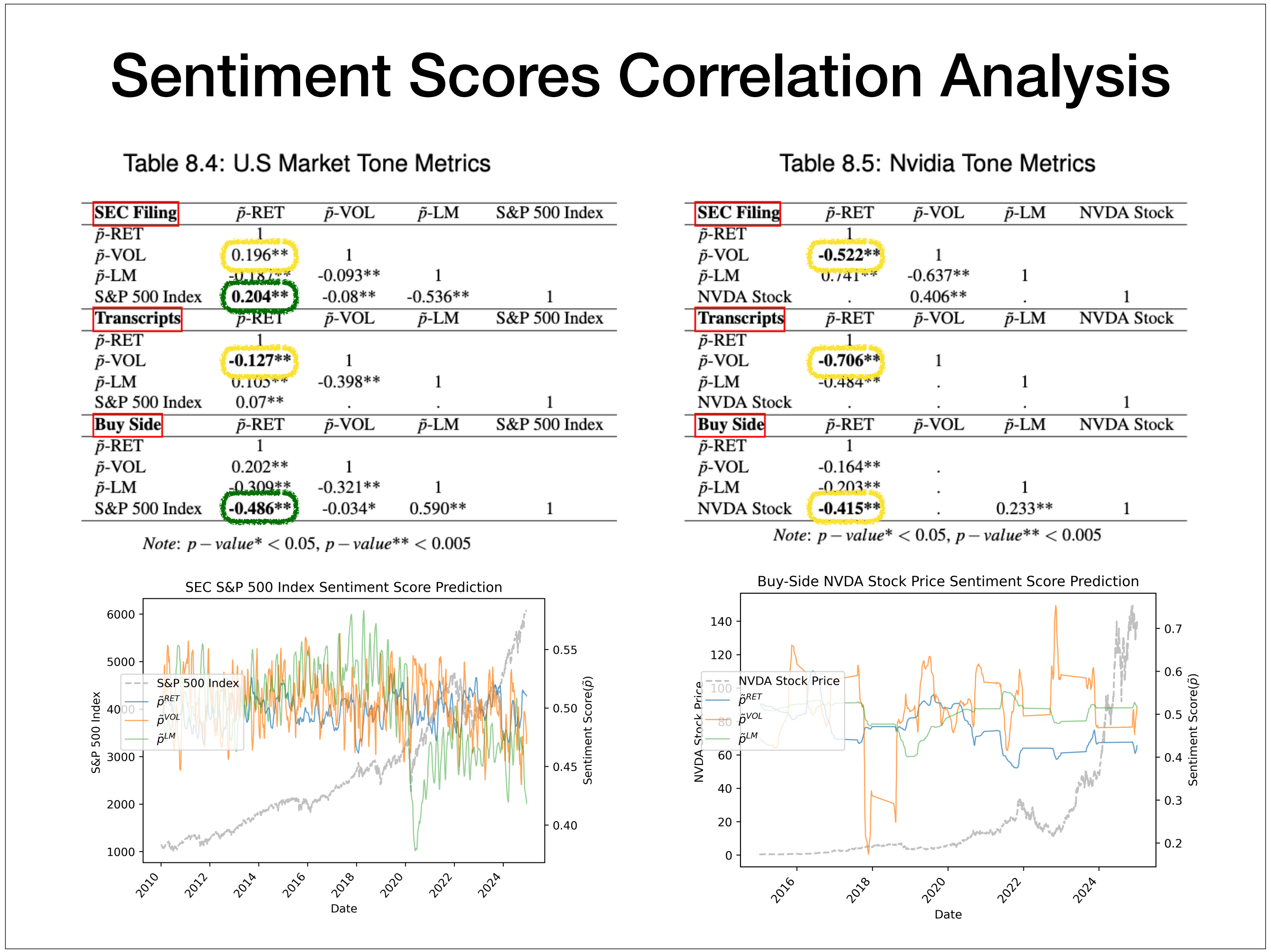
Source	Var	Mean	SD	Max/Min
SEC Filing	$\bar{p}^{RET}$	0.49	0.19	0.80/0.19
	$\bar{p}^{VOL}$	0.44	0.37	0.91/0.09
Transcript	$\bar{p}^{RET}$	0.49	0.08	0.62/0.39
	$\bar{p}^{VOL}$	0.53	0.19	0.88/0.28
Buy-Side	$\bar{p}^{RET}$	0.47	0.05	0.61/0.37
	$\bar{p}^{VOL}$	0.51	0.10	0.78/0.14

"We analyzed descriptive statistics for each sentiment signal—mean, standard deviation, and range—at both the market and firm levels.

The results show our model generates well-distributed sentiment scores, avoiding static or overly neutral outputs. It adapts its sensitivity based on context, capturing meaningful variation across data types.

Compared to our previous Minf1 work on the NASDAQ 100, where sector sentiment had a higher standard deviation around 0.24 to 0.28, the current S&P 500 market sentiment was much more stable, with SD between 0.01 and 0.03—suggesting that market sentiment serves as a reliable macro indicator.

Meanwhile, firm-level sentiment, like Nvidia's, showed much greater variability, emphasizing the model's strength in capturing company-specific tone shifts—valuable for fine-grained forecasting."



Sentiment Scores Correlation Analysis

Table 8.4: U.S Market Tone Metrics

SEC Filing	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	S&P 500 Index
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	0.196**	1		
$\tilde{p}$-LM	-0.187**	-0.093**	1	
S&P 500 Index	**0.204****	-0.08**	-0.536**	1
Transcripts	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	S&P 500 Index
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	**-0.127****	1		
$\tilde{p}$-LM	0.105**	-0.398**	1	
S&P 500 Index	0.07**	.	.	1
Buy Side	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	S&P 500 Index
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	0.202**	1		
$\tilde{p}$-LM	-0.309**	-0.321**	1	
S&P 500 Index	**-0.486****	-0.034*	0.590**	1

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

Table 8.5: Nvidia Tone Metrics

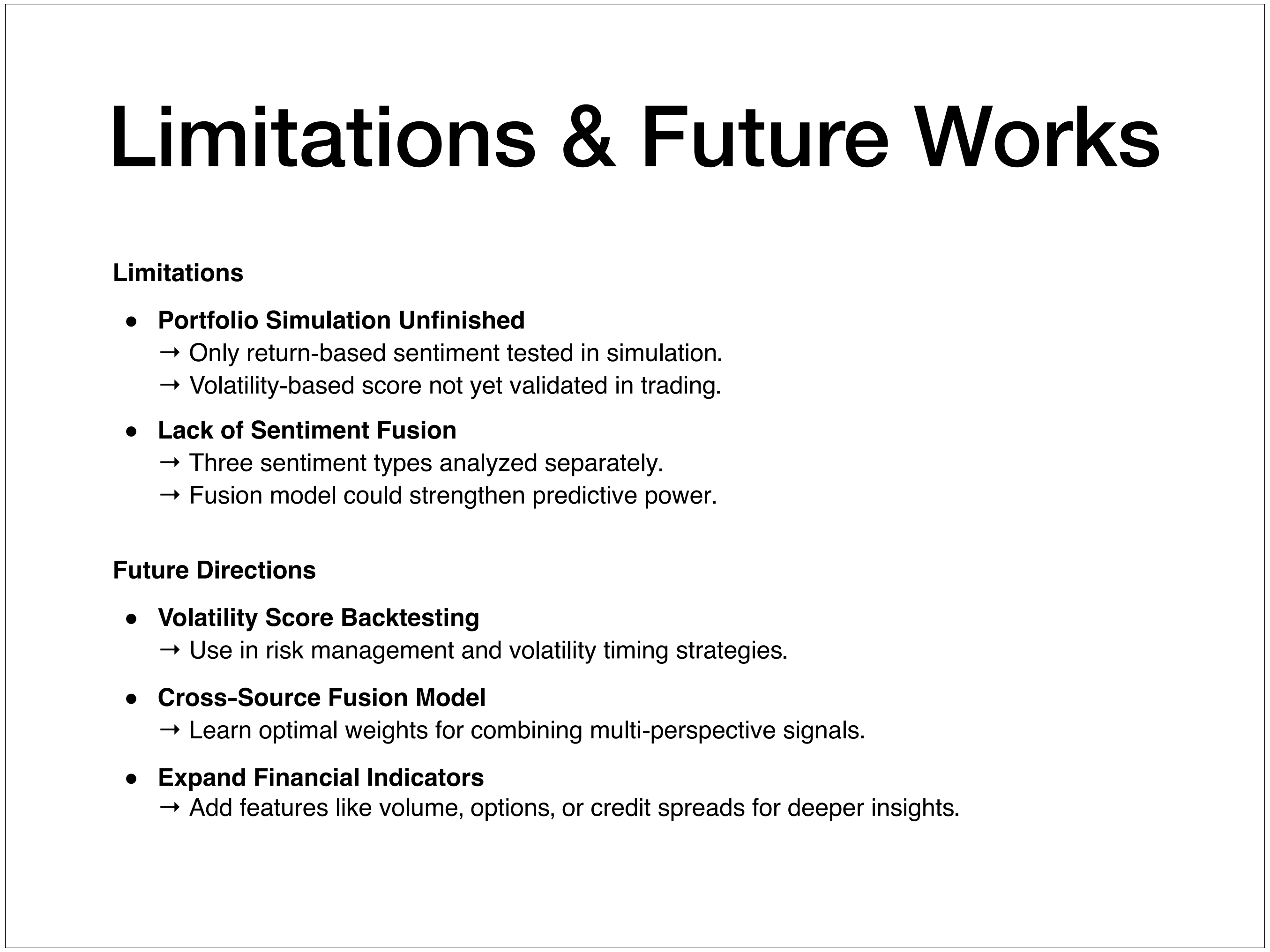
SEC Filing	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	NVDA Stock
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	**-0.522****	1		
$\tilde{p}$-LM	0.741**	-0.637**	1	
NVDA Stock	.	0.406**	.	1
Transcripts	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	NVDA Stock
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	**-0.706****	1		
$\tilde{p}$-LM	-0.484**	.	1	
NVDA Stock	.	.	.	1
Buy Side	$\tilde{p}$-RET	$\tilde{p}$-VOL	$\tilde{p}$-LM	NVDA Stock
$\tilde{p}$-RET	1			
$\tilde{p}$-VOL	-0.164**	.		
$\tilde{p}$-LM	-0.203**	.	1	
NVDA Stock	**-0.415****	.	0.233**	1

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

"Our sentiment correlation analysis revealed how different stakeholders perceive and influence market sentiment.

At the market level, sentiment from SEC filings and buy-side reports showed positive correlations between return and volatility (the first left yellow mark)—indicating that optimism often accompanies greater market risk. Interestingly, the second left yellow mark, earnings call sentiment showed an inverse relationship, suggesting that leadership's positive tone may reduce perceived risk.
Notably, SEC-trained sentiment was positively correlated with the S&P 500 index, while buy-side optimism was negatively correlated (the first second green marks), highlighting a disconnect between analyst expectations and actual market behavior.

At the firm level, Nvidia showed strong negative correlations between return and volatility sentiment across all sources, suggesting shared optimism tends to ease risk perception. Yet, buy-side optimism also correlated negatively with Nvidia's stock price—pointing to possible over-optimism or delayed market reactions."



Limitations & Future Works

Limitations

- **Portfolio Simulation Unfinished**
  → Only return-based sentiment tested in simulation.
  → Volatility-based score not yet validated in trading.
- **Lack of Sentiment Fusion**
  → Three sentiment types analyzed separately.
  → Fusion model could strengthen predictive power.

Future Directions

- **Volatility Score Backtesting**
  → Use in risk management and volatility timing strategies.
- **Cross-Source Fusion Model**
  → Learn optimal weights for combining multi-perspective signals.
- **Expand Financial Indicators**
  → Add features like volume, options, or credit spreads for deeper insights.

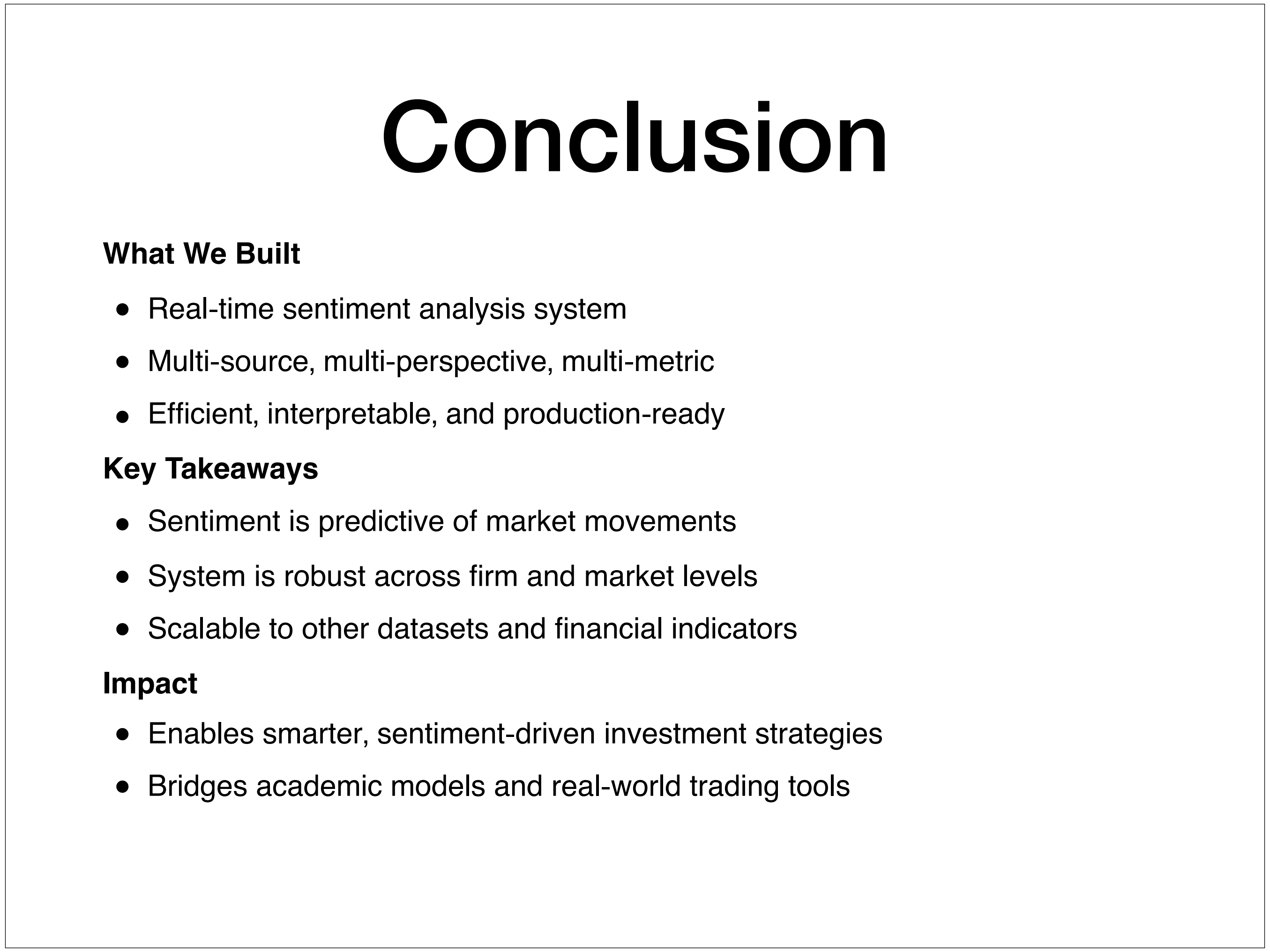
To conclude, let’s reflect on the limitations of this work and identify paths for future enhancement.

First, although we validated the predictive power of both return- and volatility-based sentiment scores, only return-based sentiment was evaluated via portfolio simulation. The volatility-based signal, while strongly correlated with actual volatility, has not been tested in a trading environment. Future work should evaluate its practical utility via simulation or backtesting.

Second, we proposed three distinct sentiment metrics based on different perspectives—company disclosures, leadership tone, and buy-side analyst views. However, we have not yet explored how to combine or synthesize these into a unified signal. Recent studies suggest that cross-source fusion could significantly boost predictive power, especially in volatile markets like crypto. A fusion model that intelligently weights these signals may further enhance forecasting accuracy.

Finally, while our current focus has been on sentiment correlation, the system could evolve toward real-time decision support for institutional trading or asset allocation. Expanding to other financial indicators—like volume, liquidity, or options data—may broaden its use case beyond return and volatility forecasting.

In summary, while our model demonstrates robust foundations, several promising research directions remain—from validating volatility-based scores in practice to building cross-source fusion and broadening financial target variables.



Conclusion

What We Built

- Real-time sentiment analysis system
- Multi-source, multi-perspective, multi-metric
- Efficient, interpretable, and production-ready

Key Takeaways

- Sentiment is predictive of market movements
- System is robust across firm and market levels
- Scalable to other datasets and financial indicators

Impact

- Enables smarter, sentiment-driven investment strategies
- Bridges academic models and real-world trading tools

To wrap up:

We developed a novel, real-time sentiment prediction system for the US stock market.

- It processes SEC filings, earnings calls, and expert reports—scalable to over 300,000 documents.
- It outputs actionable sentiment metrics that support both return and volatility forecasting.
- The system is efficient, interpretable, and extensible—built on robust engineering practices using Airflow, Docker, and Spark.

This thesis bridges the gap between theoretical sentiment modeling and real-world, real-time financial forecasting.

References

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

Thank You
Q&A