

Assessing Market Trends through Sentiment Analysis of Financial News



Group D

Hemal Mewantha	16231
Danuja Fernando	16024
Chathurangi Akmeemana	16012
Sanjana Fernando	16145

Abstract

This project designs and implements an automated cloud-based pipeline to analyze the correlation between financial news sentiment and stock price movements. The system leverages n8n for data extraction, Google Cloud Platform (GCP) for storage and processing, and Apache Airflow for orchestration to collect real-time stock data and company-specific news from Finnhub API. Through Google's Natural Language API, news headlines are scored for sentiment and analyzed using a "most impactful news" methodology to identify significant market-moving events. The results demonstrate a measurable correlation between news sentiment and stock performance, which is visualized through an interactive Looker Studio dashboard. This end-to-end solution provides investors and analysts with actionable insights, showcasing the practical application of big data technologies for financial market analysis.

Table of Contents

Abstract	1
Table of Contents	1
Table of Figures	2
1 Introduction	3
2 System Architecture and Implementation.....	4
2.1 Data Extraction	4
2.2 Data Ingestion	6
2.3 Data preprocessing.....	6
2.4 Data Orchestration with Apache Airflow.....	7
2.5 Infrastructure Management with Terraform.....	7
3 Analysis Methodology.....	8
3.1 Sentiment Analysis.....	8
3.2 Correlation Analysis.....	9
4 Dashboard Implementation.....	11
4.1 Interactive Dashboard Components.....	11
5 Challenges and Mitigation.....	14

6	Conclusion	15
7	References	16
8	Appendix	16

Table of Figures

Figure 1 - The Project Architecture.....	4
Figure 2 – n8n Data Ingestion Workflow for stock price	5
Figure 3 - n8n Data Ingestion Workflow for news headlines	5
Figure 4 - GCS bucket	6
Figure 5 - Stock Price Data Pipeline DAG	7
Figure 6 - News Sentiment Analysis Pipeline DAG	7
Figure 7 - Apache Airflow Web Interface showing the list of deployed and active DAGs.	7
Figure 8 - BigQuery Table with news headline, sentiment score and sentiment category	9
Figure 9- Weekday_stock_sentiment_correlation table in BigQuery	10
Figure 10 - Looker Studio Dashboard	11
Figure 11 - Company filter and Date range selector in dashboard	11
Figure 12 - KPI Value Boxes in Dashboard.....	12
Figure 13 - Primary Correlation Visualization in dashboard	12
Figure 14 - Trend Line of Price Difference Percentage in dashboard	13
Figure 15 - Trend Line of Sentiment Headlines in dashboard	13
Figure 16 - Sentiment Distribution of Headlines pie chart in dashboard	14
Figure 17 - News Headline Table in dashboard.....	14

1 Introduction

The stock market is a major part of the economy. It enables companies to raise money and gives investors a place to grow their wealth. The stock prices change continuously. This is due to, any things such as company earnings, economic updates, global events and most importantly, the mood of the investors. News plays a big role here because headlines and reports can quickly change how people think and act. That leads to sudden movements in stock prices.

The issue is that financial news is produced in massive amounts on a day-to-day basis and is updated in real time. While each headline may seem clear, it is difficult to measure overall pattern or compare how news affects different companies just by reading manually. This is why automated methods are helpful.

This project addresses this need by designing and implementing a comprehensive automated data pipeline. Our aim is to systematically analyze the correlation between the sentiment of financial news and subsequent stock price movements for a selection of major companies. We transform unstructured news headlines and real-time market data into clear, actionable insights.

To achieve this, the project utilizes modern technologies: n8n for data extraction from Finnhub API, Google Cloud Platform (GCP) for storage and cloud computing, Vertex AI's Natural Language API for accurate sentiment scoring and Apache Airflow for robust pipeline orchestration. The final analysis is conducted within BigQuery.

To deliver our results at once we built an interactive dashboard using Looker Studio. This dashboard illustrates graphically the correlation between stock movement and news sentiment. It allows for filtering by date and company, looking at trends over time and quick comprehension of the market's reaction to news.

The results of this project are useful to investors, traders and analysts as they provide a simple data-driven way to understand the effect of news on stock prices. It can help in making better investment choices.

2 System Architecture and Implementation

The system was developed as an end-to-end automated data pipeline that collects real-time stock data and company specific financial news, processes them for sentiment analysis and stores the results for developing the dashboard and further analytics. The architecture of our project is shown below.

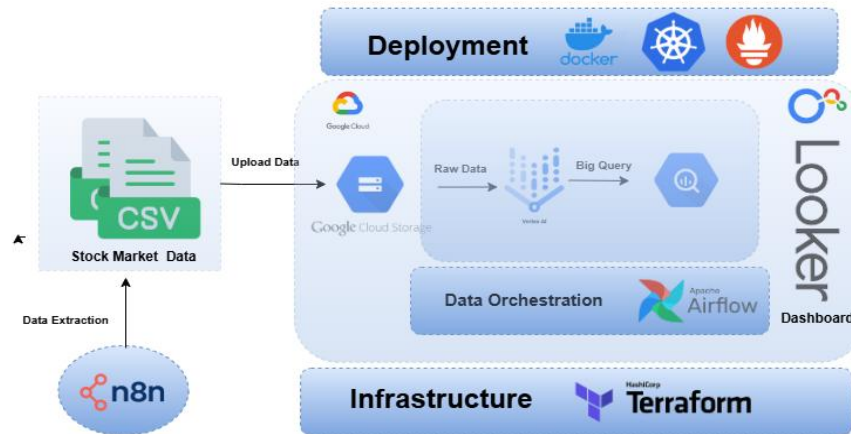


Figure 1 - The Project Architecture

2.1 Data Extraction

To facilitate efficient and automated data ingestion, a robust workflow was implemented using the n8n automation platform. This layer is responsible for the daily collection of raw stock and news data from the Finnhub API and its immediate preparation for cloud processing. The workflow is configured to be executed daily at **6:00 AM** via n8n's built-in Scheduler trigger.

2.1.1 Data Retrieval Process

By leveraging scheduled triggers and HTTPS requests, real-time data is fetched from external APIs.

- **Stock Prices** - Collected the latest stock prices for 10 companies (e.g., Apple, Tesla, Meta) using the `/quote` endpoint. Captured metrics including current price, previous close, price change (%), volume, and estimated volatility.
- **Market-related news headlines** - Retrieved market-related news using the `/news?category=general` endpoint. A custom matching algorithm associated each headline with the relevant company. To operate within the free-tier API quota,

extraction was limited to 10 headlines per company. This ensures sufficient data for daily sentiment analysis without exceeding request limits.

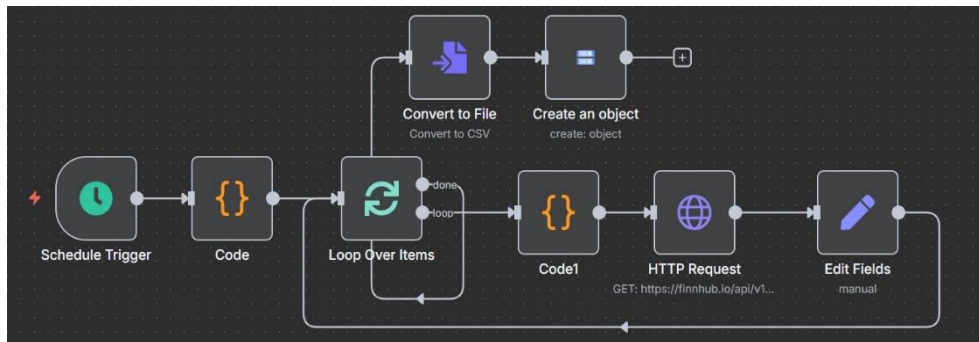


Figure 2 – n8n Data Ingestion Workflow for stock price

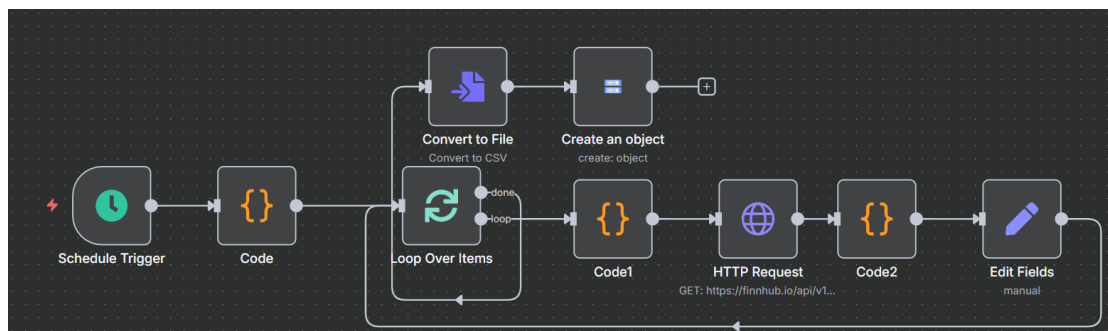


Figure 3 - n8n Data Ingestion Workflow for news headlines

2.1.2 Data Processing and Preparation

Following successful data retrieval from the Finnhub API, raw JSON responses undergo various operations in the n8n workflow to structure, clean and enrich the data, preparing it for optimal storage and analysis.

- Data structuring and Field Mapping:** The raw API output contains abbreviated field names (e.g., 'c' for current_price, 'pc' for close). A Set node is used to map these cryptic fields to human-readable, descriptive column names (e.g., current_price, previous_close), ensuring clarity and consistency for all downstream processes.
- Data Validation & Filtering:** The workflow includes logic to filter out incomplete or erroneous records and to validate data types, ensuring the integrity and quality of the data before it is persisted to the cloud.
- Format Conversion for Storage:** The final, processed JavaScript objects are passed through a **Convert to File** node, which serializes the data into CSV format. This creates

a standardized, universal file format that is optimal for storage and compatible with a wide array of data processing tools.

2.2 Data Ingestion

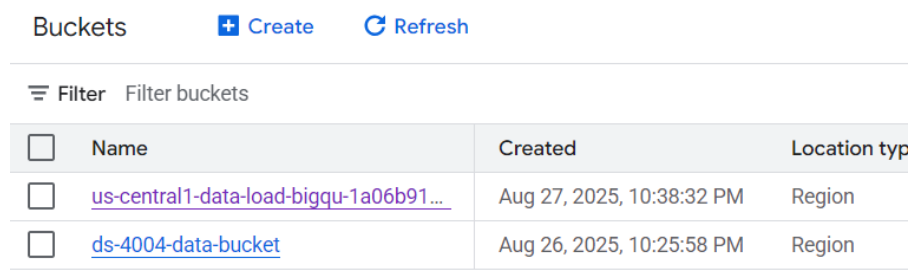
The data ingestion phase involves the automated transfer of the structured data from n8n workflows into a cloud storage environment.

2.2.1 Automated Cloud Storage Upload

The n8n workflow utilizes a dedicated Google Cloud Storage node to create a new object (file) within a predefined GCS bucket. The CSV data generated in the previous step is streamed directly to GCS.

To maintain order and ensure easy traceability, files are saved with a clear naming convention that includes the data type (news_headline, stock_price). Rather than creating a new file each day, the workflow is configured to append new daily records to the same file

This ensured that structured and queryable data is readily available in a centralized cloud environment for downstream analytics, archival, or AI-driven processing. By automating the end-to-end data flow from ingestion to cloud storage, this workflow enhances data reliability, reduces manual overhead, and supports scalable cloud-based operations.



Buckets + Create Refresh		
<input type="checkbox"/> Filter	Filter buckets	
<input type="checkbox"/> Name	Created	Location type
<input type="checkbox"/> us-central1-data-load-bigqu-1a06b91...	Aug 27, 2025, 10:38:32 PM	Region
<input type="checkbox"/> ds-4004-data-bucket	Aug 26, 2025, 10:25:58 PM	Region

Figure 4 - GCS bucket

2.3 Data preprocessing

- **Missing Value Handling:** Null or incomplete are systematically identified and removed to maintain dataset integrity.
- **Duplicate Removal:** Exact duplicate entries across all key dimensions (symbol, timestamp, headline text) are detected and eliminated to prevent analytical skew.

- **Weekend Data Filtering:** Based on financial market conventions, records timestamped on weekends are excluded from the dataset due to significantly reduced trading activity and potential data anomalies.

2.4 Data Orchestration with Apache Airflow

The data processed by GCS is Orchestrated by automated pipelines based on Apache Airflow. Two primary Directed Acyclic Graphs (DAGs) were designed for managing the workflow of stock price data and news sentiment analysis respectively.

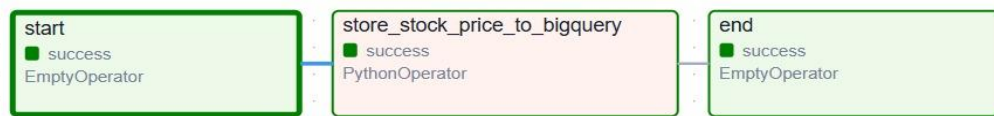


Figure 5 - Stock Price Data Pipeline DAG



Figure 6 - News Sentiment Analysis Pipeline DAG

The designed Airflow DAGs have been successfully deployed and are running in a production environment.

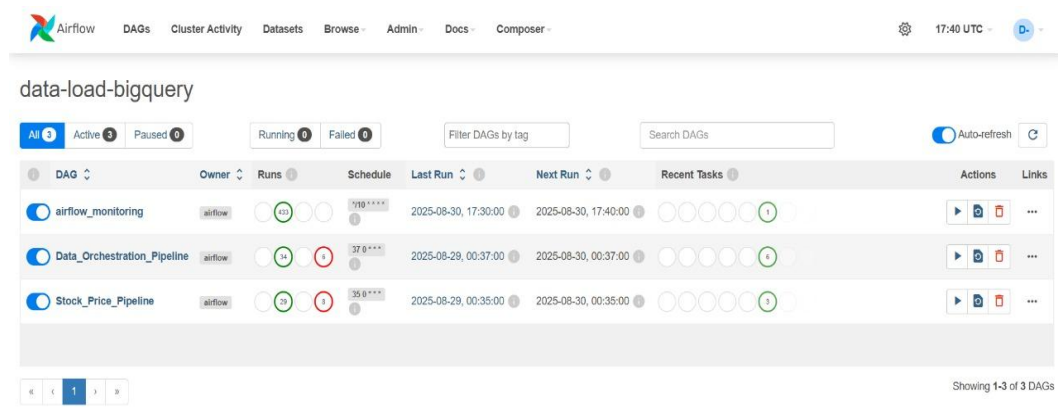


Figure 7 - Apache Airflow Web Interface showing the list of deployed and active DAGs.

2.5 Infrastructure Management with Terraform

We used **Terraform**, a tool that lets us set up and manage our Google Cloud resources using code instead of doing it manually. With simple configuration files, we automatically created:

- **Storage buckets** to safely hold all the stock and news data
- **BigQuery datasets** to organize and analyze the information efficiently

This approach made sure our system was consistent, easy to reproduce, and could grow as needed. It also reduced errors and saved time by automating the setup process. Terraform helped us build a strong, reliable foundation for our entire project.

3 Analysis Methodology

3.1 Sentiment Analysis

The core of our news pipeline processing is extracting quantitative sentiment scores from text news headlines. Google Cloud's Natural Language API (which is a pre-trained machine learning model for natural language understanding) used to transform unstructured text to structured analytical data.

3.1.1 Sentiment Scoring Process

The sentiment analysis is executed within the `vertex_ai_sentiment` task of our Airflow DAG. The process for each news headline is as follows:

- **API Request:** The headline is sent to the Google Cloud Natural Language API with the `'analyze_sentiment'` method. The API examines the emotional tone of text.
- **Score Extraction:** The API returns a `'sentiment_score'` ranging between -1.0 (extremely negative) and 1.0 (extremely positive). This real-valued score is a fine-grained estimate of the emotional polarity expressed in the headline.
- **Rate Limiting:** 100-millisecond delay (`'time.sleep(0.1)'`) is inserted between API calls. This is a critical implementation detail to respect the API's rate limits on the free plan to prevent quota errors and ensure stable, uninterrupted operation.

3.1.2 Sentiment Categorization

To make it easier to understand and allow subsequent analysis and dashboarding, continuous sentiment scores are converted to discrete categories in the `'categorize_sentiment'` task:

- Positive - `'sentiment_score > 0.1'`
- Neutral - `'-0.1 <= sentiment_score <= 0.1'`
- Negative - `'sentiment_score < -0.1'`

3.1.3 Data Enrichment and Storage

The last output of the sentiment analysis pipeline is an enriched dataset that combines the raw news metadata with calculated sentiment metrics. Each record contains:

- The company symbol
- The headline text
- The original date fields
- Sentiment_score
- Sentiment Category

The data set is automatically appended to a partitioned table in BigQuery ('ds_4004_news_headline') to be used for correlation analysis with day-to-day stock price fluctuation as well as for visualization in dashboards.

Schema	Details	Preview	Table Explorer	Preview	Insights	Lineage	Data Profile	Data Quality
Row	from_date	to_date	company	headline	sentiment_s...	sentiment_category		
1	2025-08-16	2025-08-17	AMZN	Jim Cramer Says Don't Quit Market When It's Frothy: 'Is Widespread Irrationality a Reason To Sell Down in Perfectly Rational Stocks?'	-0.60000002...	Negative		
2	2025-08-16	2025-08-17	AAPL	Meta spends more guarding Mark Zuckerberg than Apple, Nvidia, Microsoft, Amazon, and Alphabet do for their own CEOsâ€"combined	-0.30000001...	Negative		
3	2025-08-16	2025-08-17	GOOGL	Meta spends more guarding Mark Zuckerberg than Apple, Nvidia, Microsoft, Amazon, and Alphabet do for their own CEOsâ€"combined	-0.30000001...	Negative		
4	2025-08-16	2025-08-17	AMZN	Meta spends more guarding Mark Zuckerberg than Apple, Nvidia, Microsoft, Amazon,	-0.30000001...	Negative		

Figure 8 - BigQuery Table with news headline, sentiment score and sentiment category

3.2 Correlation Analysis

After saving sentiment-enriched news data and stock prices in BigQuery, the next task was to quantitatively measure the relationship between them.

3.2.1 Sentiment Aggregation

One of the essential tasks was aggregating multiple daily sentiment scores into a single value representative of the day's news impact for each company. Two primary methods were under consideration:

1. Simple Average: The initial method calculated the average of the overall sentiment scores of a company on a particular day. This was not utilized due to the possibility of cancelling each other out by highly negative and highly positive headlines that neutralized one another. It dilute the true market signal and also compusing overall sentiment of the day.
2. Most impactful News: For each company on each day, the headline with absolute maximum sentiment score was selected. This number, being the most emotionally charged news of the day and it was used for correlation with the daily stock price change.

3.2.2 Correlation Implementation

Correlation analysis was also performed successfully in BigQuery through a SQL query. The process first merges stock price data with the most impactful news headlines of each company on respective dates. It then calculates the correlation coefficient between daily sentiment score and the price changes for each company using BigQuery's built-in CORR function. This correlation value represents the strength and direction of the linear relationship between that company's daily "most impactful" sentiment scores and its daily stock price difference percentage over the entire analysis period. The final output is a table that integrates the raw

Schema	Details	Preview	Table Explorer	Preview	Insights	Lineage	Data Profile	Data Quality
Row	company	trade_day	sentiment_s...	headline	current_price	price_diff	diff_percenta...	corr_sentime...
1	CRM	2025-08-18	-0.80000001...	HR giant Workday says hackers stole personal data in recent breach	243.97	1.53	0.6311	0.027779169...
2	AMZN	2025-08-18	-0.69999998...	Amazon accused of driving up prices for Britons in class action lawsuit	231.49	0.46	0.1991	0.150321670...
3	META	2025-08-18	-0.69999998...	Meta's reported shake-up, Hims & Hers sinks on GoodRx's Novo deal	767.37	-17.86	-2.2745	-0.03516264...
4	NVDA	2025-08-18	-0.5	AMD: Market Says Time To Get ...	182.01	1.56	0.8645	0.257019019...
5	NFLX	2025-08-18	-0.5	Should You Invest in Netflix (NFLX) Based on Bullish Wall Street Views?	1245.09	6.14	0.4956	0.246960637...
6	GOOGL	2025-08-18	-0.30000001...	Google Expands Stake In Data Center Firm Whose Stock	203.5	-0.4	-0.1962	-0.16290097...

Figure 9- Weekday_stock_sentiment_correlation table in BigQuery

data with the calculated correlation coefficients. Now it is ready for visualization and further analysis.

4 Dashboard Implementation

An interactive Google Looker Studio dashboard was created to visualize and analyze correlations between daily market sentiment and stock price movements. The dashboard connects directly to BigQuery tables updated by the automated sentiment analysis pipeline, providing a dynamic interface for real-time insights. This visualization layer enables stakeholders to intuitively explore correlations between news sentiment and stock performance.



Figure 10 - Looker Studio Dashboard

4.1 Interactive Dashboard Components

4.1.1 Control Mechanisms



Figure 11 - Company filter and Date range selector in dashboard

The dashboard is user-interactive and includes two primary filters for targeted analysis, with a Company Filter for selecting individual companies and a Date Range Selector for customizing the analysis period.

4.1.2 KPI value Boxes

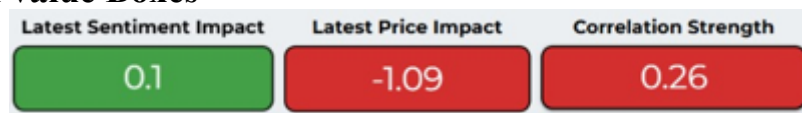


Figure 12 - KPI Value Boxes in Dashboard

4.1.2.1 Latest Sentiment Impact

Shows the strongest positive or negative news sentiment for the latest trading day, indicating whether recent headlines were favorable, neutral, or adverse.

4.1.2.2 Latest Price Impact

Displays the stock's percentage change for the latest day, showing how the market reacted.

4.1.2.3 Correlation Strength

Measures the historical relationship between sentiment and price movement, indicating how reliably news sentiment explains stock behavior.

4.1.3 Primary Correlation Visualization

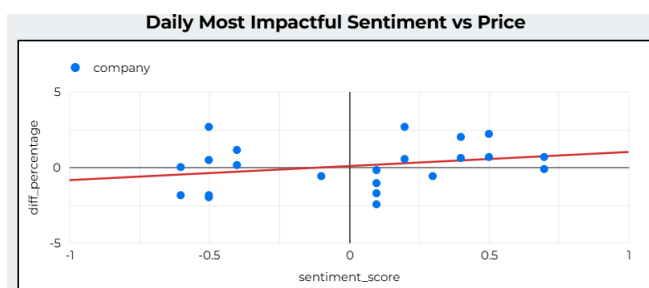


Figure 13 - Primary Correlation Visualization in dashboard

The visualization illustrates the relationship between news sentiment scores and daily stock price changes. Each point represents the link between sentiment score and price difference percentage for a particular company within a given date range, while the trend line highlights the overall correlation.

4.1.4 Trend Line of Price Difference Percentage

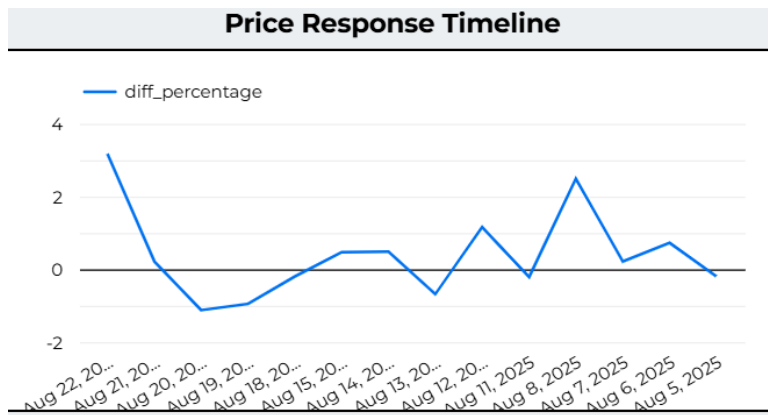


Figure 14 - Trend Line of Price Difference Percentage in dashboard

The visualization tracks stock price difference percentages over time for each company within a given date range. It emphasizes how price movements evolve day by day, allowing the identification of patterns, fluctuations, and potential trigger points in response to market events or sentiment shifts.

4.1.5 Trend Line of Sentiment Headlines

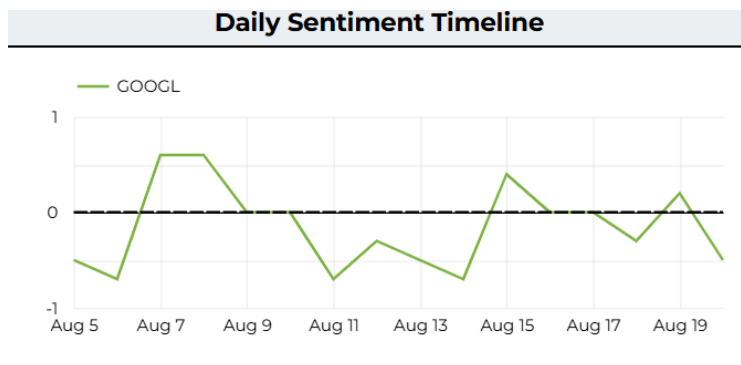


Figure 15 - Trend Line of Sentiment Headlines in dashboard

The visualization presents sentiment trends over time for each company within a given date range. It captures how daily news sentiment fluctuates and highlights its potential influence on stock price differences, offering insights into the temporal relationship between market perception and price response.

4.1.6 Sentiment Distribution of Headlines

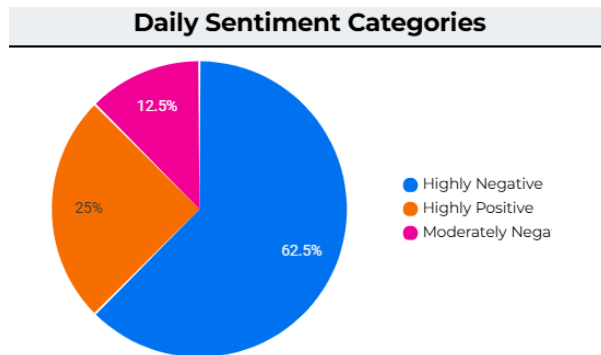


Figure 16 - Sentiment Distribution of Headlines pie chart in dashboard

The visualization illustrates how sentiment evolves for each company within a given date range. It highlights the distribution of positive, negative, and neutral news headlines received daily, providing insights into overall market perception and its potential connection to stock price movements.

4.1.7 News Headlines Table

Most Impactful Headlines Table				
	trade_day	headline	sentiment_score	diff_percentage
8.	Aug 28, 2025	US FTC chair alleges Gmail uses 'partisan filtering'	-0.60	2.005
9.	Aug 8, 2025	TY: Large Discount And Decent Performance From This CEF	0.60	2.4934
10.	Aug 13, 2025	CoreWeave stock plummets as AI cloud company reports 'deteriorating' operating income outlook	-0.50	-0.6787
11.	Aug 20, 2025	Thousands of Grok chats are now searchable on Google	-0.50	-1.1162
12.	Aug 5, 2025	I Bought Amazon As I'm Ultra Bullish On Its Margin-Heavy Era Underway	-0.50	-0.1897
13.	Aug 25, 2025	NotebookLM's Video Overviews feature now supports 80 languages	0.50	1.1645

Figure 17 - News Headline Table in dashboard

This table presents company-specific news headlines ordered by their sentiment impact. Headlines with the strongest positive or negative sentiment appear at the top, enabling users to quickly identify the most influential news of the day. Each entry includes the sentiment score, news category, and publication date, providing a clear view of which developments are driving market sentiment.

5 Challenges and Mitigation

I. API Rate Limiting and Quota Management:

- **Challenge:** The free version of the Finnhub API only allows a limited number of requests per minute. Our automated workflows could easily send too many requests too quickly and get blocked.

- **Solution:** We added short pauses between our API calls to slow down the requests. We also limit ourselves to 10 news articles per company to stay under the daily limit.

II. Data Quality and Noise:

- **Challenge:** Raw data contained inconsistencies, missing values, duplicates, and non-trading day (weekend) records, which could skew analysis.
- **Solution:** Built robust preprocessing pipelines in n8n and Airflow to validate, clean, and filter data before storage. Removed weekends and duplicates rigorously.

III. Combining Daily Sentiment scores

- **Challenge:** Each day, we had 10 sentiment scores for each company. Using a simple average was a bad idea because a very positive and a very negative score would cancel each other out, making the result look "neutral" and hiding the true effect of the news.
- **Solution:** We decided to use only the **most extreme score** each day (the most positive OR the most negative headline). This gave us a clearer picture of the biggest news impact on the company each day.

6 Conclusion

This project successfully developed and implemented a complete automated pipeline for analyzing the sentiment of financial news and stock price change. By leveraging modern cloud technologies including n8n, Google Cloud Platform, Apache Airflow and Looker Studio, we constructed a scalable system that transforms unstructured news data into actionable investment insights.

The Key achievements include:

- Automating collection and cleaning of stock prices and news headlines from reliable sources.
- Effective use of cloud-based NLP to generate accurate sentiment scores.
- Establishing a solid methodology to identify and analyze daily impactful sentiment.
- Demonstrating measurable correlation between news sentiment and stock price movement.

- Creating an intuitive dashboard that makes complex analytical results accessible to non-technical users.
- Building a robust, automated pipeline that requires minimal manual intervention.

This project fulfills its objectives of creating a practical, data-driven tool for financial market analysis while demonstrating the effective application of big data technologies in the financial sector.

7 References

Build infrastructure | Terraform | HashiCorp Developer. (n.d.). Retrieved July 15, 2025, from <https://developer.hashicorp.com/terraform/tutorials/gcp-get-started/google-cloud-platform-build>

Docs overview | hashicorp/google | Terraform | Terraform Registry. (n.d.). Retrieved August 8, 2025, from <https://registry.terraform.io/providers/hashicorp/google/latest/docs>

GCP Composer | Airflow GCS to BigQuery and BigQuery Operators - YouTube. (n.d.). Retrieved July 15, 2025, from <https://www.youtube.com/watch?v=OBRv3t697sQ&list=PL7B3mwEXCi-aLpjswSYJkaiCehxQ2Xz39&index=3>

Google Cloud Provider Configuration Reference | Guides | hashicorp/google | Terraform | Terraform Registry. (n.d.). Retrieved July 15, 2025, from https://registry.terraform.io/providers/hashicorp/google/latest/docs/guides/provider_reference

Terraform overview | Terraform | HashiCorp Developer. (n.d.). Retrieved August 8, 2025, from <https://developer.hashicorp.com/terraform/docs>

8 Appendix

Data and Infrastructure Integration: <https://github.com/hemalmewan/Group-D-Big-Data-Project.git>