Predicting Nvidia Stock Movements Using News Headlines and Sentiment Analysis

Abstract

This project builds an end-to-end pipeline to study how news headlines about Nvidia relate to short-term movements in Nvidia's stock price. I first use the Google News feed to scrape around two years of Nvidia-related headlines, storing the source, headline text, and publication date in a structured CSV file. I then download matching daily stock price data for Nvidia (ticker NVDA) from Yahoo Finance, including open, close, high, low, and volume. After cleaning and merging the data by date, I engineer features such as daily average sentiment score and news count, and define a binary target indicating whether the next day's closing price goes up or not. Finally, I train and compare two classification models, Random Forest and XGBoost, using a time-series cross-validation scheme and evaluate them using accuracy, precision, recall, F1-score, and ROC-AUC. The results illustrate that combining basic sentiment information from headlines with price and volume can provide some predictive signal, but also highlight limitations and opportunities for improving the modelling approach.

---

1. Introduction

Financial markets react quickly to new information, and company-specific news is one of the most visible sources of such information. For large technology firms like Nvidia, news about product launches, earnings, regulation, or macroeconomic conditions can influence investor expectations and short-term stock price movements. At the same time, recent advances in natural language processing and machine learning make it easier to quantify textual information such as headlines and integrate it with traditional numerical market data.

The main aim of this project is to investigate whether daily news headlines about Nvidia contain useful information for predicting whether the Nvidia stock price will move up or down on the next trading day. To answer this question, I construct a dataset that combines scraped Google News headlines for Nvidia with historical daily stock prices for the NVDA ticker. I then derive simple sentiment and volume features from the news data and train supervised classification models to predict the direction of the next day's closing price.

The specific objectives of the project are:

- To collect approximately two years of Nvidia-related news headlines from Google News and store them in a structured format.
- To download matching NVDA daily stock prices from Yahoo Finance and prepare a clean price dataset.
- To perform basic data cleaning, date handling, and feature engineering, including headline length, sentiment, and news counts per day.
- To merge the daily news features with stock prices and define a binary target for next-day price direction.
- To train and compare Random Forest and XGBoost classifiers using time-series cross-validation and standard evaluation metrics.

By the end of the report, I discuss how well these models perform, what their limitations are, and how the pipeline could be extended or improved in future work.

---

2. Data Collection

2.1 News scraping with Google News

The first part of the project focuses on collecting news headlines related to Nvidia from the Google News feed. I implement a function fetch_nvidia_headlines_clean() using the pygooglenews library to search for the keyword “Nvidia” over a rolling window of monthly time ranges covering roughly the last two years. Inside a loop, the function moves backwards one month at a time, constructing from and to dates, and calling gn.search('Nvidia', from_=start_str, to_=end_str) to retrieve the entries for that period.

For each entry in the search results, I extract three key fields: the news source name, the headline text, and the publication date. These are stored in a list of dictionaries and then converted into a pandas DataFrame. I keep only the relevant columns source, headline, and pubdate, and drop duplicate rows based on the headline text to avoid repeated entries. Finally, the cleaned DataFrame is saved as news_raw.csv, and the function prints how many headlines were collected; in this run, 2328 unique headlines were stored.

	source	headline	pubdate
1	NVIDIA Newsroom	NVIDIA and CoreWeave Strengthen Collaboration to Accelerate Buildout of AI Factories - NVIDIA Newsroom	Mon, 26 Jan 2026 13:30:39 GMT
2	NVIDIA Newsroom	NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer - NVIDIA Newsroom	Mon, 05 Jan 2026 08:00:00 GMT
3	NVIDIA	NVIDIA App Update Adds DLSS 4.5 Super Resolution & New GeForce Game Ready Driver - NVIDIA	Mon, 05 Jan 2026 08:00:00 GMT
4	NVIDIA Newsroom	NVIDIA Releases New Physical AI Models as Global Partners Unveil Next-Generation Robots - NVIDIA Newsroom	Mon, 05 Jan 2026 08:00:00 GMT
5	NVIDIA Blog	NVIDIA Rubin Platform, Open Models, Autonomous Driving: NVIDIA Presents Blueprint for the Future at CES - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
6	NVIDIA Newsroom	Siemens and NVIDIA Expand Partnership to Build the Industrial AI Operating System - NVIDIA Newsroom	Tue, 06 Jan 2026 08:00:00 GMT
7	NVIDIA Blog	NVIDIA Launches Earth-2 Family of Open Models — the World's First Fully Open Set of Models and Tools for AI Weather - NVIDIA Blog	Mon, 26 Jan 2026 14:03:04 GMT
8	NVIDIA Developer	Inside the NVIDIA Rubin Platform: Six New Chips, One AI Supercomputer | NVIDIA Technical Blog - NVIDIA Developer	Mon, 05 Jan 2026 08:00:00 GMT
9	NVIDIA	NVIDIA DLSS 4.5 Delivers Major Upgrade With 2nd Gen Transformer Model For Super Resolution & 6X Dynamic Multi Frame Generation - NVIDIA	Tue, 06 Jan 2026 08:00:00 GMT
10	NVIDIA Newsroom	NVIDIA BioNeMo Platform Adopted by Life Sciences Leaders to Accelerate AI-Driven Drug Discovery - NVIDIA Newsroom	Mon, 12 Jan 2026 08:00:00 GMT
11	NVIDIA Blog	NVIDIA Brings GeForce RTX Gaming to More Devices With New GeForce NOW Apps for Linux PC and Amazon Fire TV - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
12	NVIDIA Blog	NVIDIA DGX SuperPOD Sets the Stage for Rubin-Based Systems - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
13	NVIDIA Newsroom	NVIDIA Announces Alpamayo Family of Open-Source AI Models and Tools to Accelerate Safe, Reasoning-Based Autonomous Vehicle Development - NVIDIA Newsroom	Mon, 05 Jan 2026 08:00:00 GMT
14	NVIDIA Blog	Steel, Sensors and Silicon: How Caterpillar Is Bringing Edge AI to the Jobsite - NVIDIA Blog	Wed, 07 Jan 2026 08:00:00 GMT
15	NVIDIA Blog	CEOs of NVIDIA and Lilly Share 'Blueprint for What Is Possible' in AI and Drug Discovery - NVIDIA Blog	Tue, 13 Jan 2026 08:00:00 GMT
16	NVIDIA Blog	NVIDIA RTX Accelerates 4K AI Video Generation on PC With LTX-2 and ComfyUI Upgrades - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
17	NVIDIA Newsroom	NVIDIA BlueField-4 Powers New Class of AI-Native Storage Infrastructure for the Next Frontier of AI - NVIDIA Newsroom	Mon, 05 Jan 2026 08:00:00 GMT
18	NVIDIA Blog	NVIDIA DLSS 4.5, Path Tracing and G-SYNC Pulsar Supercharge Gameplay With Enhanced Performance and Visuals - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
19	NVIDIA Blog	GeForce NOW Rings In 2026 With 14 New Games in January - NVIDIA Blog	Thu, 01 Jan 2026 08:00:00 GMT
20	NVIDIA Blog	NVIDIA Expands Global DRIVE Hyperion Ecosystem to Accelerate the Road to Full Autonomy - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT
21	NVIDIA Blog	NVIDIA DRIVE AV Software Debuts in All-New Mercedes-Benz CLA - NVIDIA Blog	Mon, 05 Jan 2026 08:00:00 GMT

2.2 Stock price data from Yahoo Finance

To complement the news data, I need the corresponding stock price information for Nvidia. Using the yfinance library, I call yf.download('NVDA', period='2y') to download about two years of daily stock prices for the NVDA ticker. From this data, I select the Open, High, Low, Close, and Volume columns, which provide the key price and trading activity variables for each day. Sometimes yfinance returns a DataFrame with multi-level columns, so I flatten the column index to keep only the top-level names.

Next, I reset the index so that Date becomes an explicit column instead of an index. The final stock DataFrame, containing one row per trading day with its prices and volume, is saved to stock_data.csv. This stock dataset will later be merged with the daily news features by date to build the modelling dataset.

Date	Open	High	Low	Close	Volume
2024-01-29	61.197906346142304	62.45420645662032	60.87309034622641	62.43022155761719	348733000
2024-01-30	62.86497347083293	63.45764174361534	62.22532667454649	62.73904037475586	410735000
2024-01-31	61.40579108468123	62.23433190687522	60.666205224632265	61.49274444580078	453795000
2024-02-01	62.065426957341195	63.15582300584299	61.61568052564377	62.991912841796875	369146000
2024-02-02	63.93837399186544	66.56291110918288	63.654531745686434	66.12316131591797	476578000
2024-02-05	68.18701012748092	69.45830503749701	67.16758142852466	69.29339599609375	680078000
2024-02-06	69.59122707783612	69.71515858330277	66.26308702618283	68.18501281738281	683111000
2024-02-07	68.28095938687078	70.18090186277372	67.56235796585537	70.05996704101562	495575000
2024-02-08	70.0349761659854	70.75457644757603	69.41632519274872	69.60221862792969	414422000
2024-02-09	70.49373428612587	72.14481511205457	70.17291382776136	72.09284973144531	436637000
2024-02-12	72.55957267433736	74.56945447682682	71.21032591879988	72.20777130126953	613710000
2024-02-13	70.36081744123659	73.40911526024477	69.58125280309775	72.08785247802734	602580000

---

3. Understanding the XML / RSS Structure

The news headlines accessed via pygooglenews originate from feeds that typically follow RSS or Atom formats. These feeds are structured XML documents designed to syndicate content from news websites and blogs in a machine-readable way. In a typical RSS feed, high-level information about the feed appears within a channel tag, while individual news entries are represented by a sequence of item elements.

Each item element normally contains fields such as title (the headline), link (the URL to the full article), and pubDate (the publication date). In the notebook, I note that each item corresponds to a single article, blog post, update, or announcement in the feed. When rendering on the website, the title is frequently displayed inside an HTML h3 tag, which is one of the ways to visually identify the article headline. Understanding this structure is useful because it clarifies how tools like pygooglenews are able to parse and expose the relevant fields that I later store in the news_raw.csv file.

---

4. Data Cleaning and Feature Engineering

4.1 Cleaning news data

The raw news CSV contains the news source, headline text, and publication date as they were retrieved from Google News. To make this data easier to work with, I perform a series of cleaning steps in pandas. First, I convert the pubdate column to a proper date type using pd.to_datetime(df["pubdate"], utc=True).dt.date, which standardises the format and strips the time component. This ensures that dates will align correctly when I later join the news data with daily stock prices.

I also create a derived feature called headline_length by converting each headline to a string and computing its character length with astype(str).apply(len). This feature captures how long or short a headline is and can potentially be used as a proxy for the complexity or richness of the news item. After these transformations, the cleaned DataFrame is saved as news_cleaned.csv for subsequent analysis.

	source	headline	pubdate	headline_length
2	NVIDIA Newsroom	NVIDIA and CoreWeave Strengthen Collaboration to Accelerate Buildout of AI Factories - NVIDIA Newsroom	2026-01-26	102
3	NVIDIA Newsroom	NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer - NVIDIA Newsroom	2026-01-05	120
4	NVIDIA	NVIDIA App Update Adds DLSS 4.5 Super Resolution & New GeForce Game Ready Driver - NVIDIA	2026-01-05	89
5	NVIDIA Newsroom	NVIDIA Releases New Physical AI Models as Global Partners Unveil Next-Generation Robots - NVIDIA Newsroom	2026-01-05	105
6	NVIDIA Blog	NVIDIA Rubin Platform, Open Models, Autonomous Driving: NVIDIA Presents Blueprint for the Future at CES - NVIDIA Blog	2026-01-05	117
7	NVIDIA Newsroom	Siemens and NVIDIA Expand Partnership to Build the Industrial AI Operating System - NVIDIA Newsroom	2026-01-06	99
8	NVIDIA Blog	NVIDIA Launches Earth-2 Family of Open Models — the World's First Fully Open Set of Models and Tools for AI Weather - NVIDIA Blog	2026-01-26	129
9	NVIDIA Developer	Inside the NVIDIA Rubin Platform: Six New Chips, One AI Supercomputer | NVIDIA Technical Blog - NVIDIA Developer	2026-01-05	112
10	NVIDIA	NVIDIA DLSS 4.5 Delivers Major Upgrade With 2nd Gen Transformer Model For Super Resolution & 6X Dynamic Multi Frame Generation - NVIDIA	2026-01-06	135
11	NVIDIA Newsroom	NVIDIA BioNeMo Platform Adopted by Life Sciences Leaders to Accelerate AI-Driven Drug Discovery - NVIDIA Newsroom	2026-01-12	113
12	NVIDIA Blog	NVIDIA Brings GeForce RTX Gaming to More Devices With New GeForce NOW Apps for Linux PC and Amazon Fire TV - NVIDIA Blog	2026-01-05	120
13	NVIDIA Blog	NVIDIA DGX SuperPOD Sets the Stage for Rubin-Based Systems - NVIDIA Blog	2026-01-05	72
14	NVIDIA Newsroom	NVIDIA Announces Alpamayo Family of Open-Source AI Models and Tools to Accelerate Safe, Reasoning-Based Autonomous Vehicle Development - NVIDIA Newsroom	2026-01-05	152
15	NVIDIA Blog	Steel, Sensors and Silicon: How Caterpillar Is Bringing Edge AI to the Jobsite - NVIDIA Blog	2026-01-07	92
16	NVIDIA Blog	CEOs of NVIDIA and Lilly Share 'Blueprint for What Is Possible' in AI and Drug Discovery - NVIDIA Blog	2026-01-13	102
17	NVIDIA Blog	NVIDIA RTX Accelerates 4K AI Video Generation on PC With LTX-2 and ComfyUI Upgrades - NVIDIA Blog	2026-01-05	97
18	NVIDIA Newsroom	NVIDIA BlueField-4 Powers New Class of AI-Native Storage Infrastructure for the Next Frontier of AI - NVIDIA Newsroom	2026-01-05	117
19	NVIDIA Blog	NVIDIA DLSS 4.5, Path Tracing and G-SYNC Pulsar Supercharge Gameplay With Enhanced Performance and Visuals - NVIDIA Blog	2026-01-05	120
20	NVIDIA Blog	GeForce NOW Rings In 2026 With 14 New Games in January - NVIDIA Blog	2026-01-01	68
21	NVIDIA Blog	NVIDIA Expands Global DRIVE Hyperion Ecosystem to Accelerate the Road to Full Autonomy - NVIDIA Blog	2026-01-05	100
22	NVIDIA Blog	NVIDIA DRIVE AV Software Debuts in All-New Mercedes-Benz CLA - NVIDIA Blog	2026-01-05	74

4.2 Identifying non-trading days

News is published every day, including weekends and holidays, whereas the stock market only trades on specific business days. To distinguish between trading and non-trading days in the news data, I merge the cleaned news DataFrame with the stock price DataFrame on the date. In code, both pubdate and Date are converted to datetime, and a left join is performed with pd.merge to combine them.

After the merge, each news row has the associated stock columns if the date is a trading day, and NaN values if there was no trading on that date. I define a boolean flag is_trading_day as Open.notna(), which is True when the market was open and False otherwise. The merged dataset, including this indicator, is saved as merged_midterm_data.csv. In a short written answer in the notebook, I also note example dates that appear as non-trading days and mention that weekends are a common reason for markets being closed.

	source	headline	pubdate	headline_length	Date	Open	High	Low	Close	Volume	is_trading_day
2	NVIDIA Newsroom	NVIDIA and CoreWeave Strengthen Collaboration to Accelerate Buildout of AI Factories - NVIDIA Newsroom	2026-01-26	102	2026-01-26	187.16000366210935	188.1199951171875	185.9900054931641	186.47000122070312	12479660.0	True
3	NVIDIA Newsroom	NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer - NVIDIA Newsroom	2026-01-05	120	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
4	NVIDIA	NVIDIA App Update Adds DLSS 4.5 Super Resolution & New GeForce Game Ready Driver - NVIDIA	2026-01-05	89	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
5	NVIDIA Newsroom	NVIDIA Releases New Physical AI Models as Global Partners Unveil Next-Generation Robots - NVIDIA Newsroom	2026-01-05	105	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
6	NVIDIA Blog	NVIDIA Rubin Platform, Open Models, Autonomous Driving: NVIDIA Presents Blueprint for the Future at CES - NVIDIA Blog	2026-01-05	117	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
7	NVIDIA Newsroom	Siemens and NVIDIA Expand Partnership to Build the Industrial AI Operating System - NVIDIA Newsroom	2026-01-06	99	2026-01-06	190.52000427246094	192.16999816894531	186.82000732421875	187.24000549316406	17686260.0	True
8	NVIDIA Blog	NVIDIA Launches Earth-2 Family of Open Models — the World's First Fully Open Set of Models and Tools for AI Weather - NVIDIA Blog	2026-01-26	129	2026-01-26	187.16000366210935	188.1199951171875	185.9900054931641	186.47000122070312	12479660.0	True
9	NVIDIA Developer	Inside the NVIDIA Rubin Platform: Six New Chips, One AI Supercomputer	NVIDIA Technical Blog - NVIDIA Developer	2026-01-05	112	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
10	NVIDIA	NVIDIA DLSS 4.5 Delivers Major Upgrade With 2nd Gen Transformer Model For Super Resolution & 6X Dynamic Multi Frame Generation - NVIDIA	2026-01-06	135	2026-01-06	190.52000427246094	192.16999816894531	186.82000732421875	187.24000549316406	17686260.0	True
11	NVIDIA Newsroom	NVIDIA BioNeMo Platform Adopted by Life Sciences Leaders to Accelerate AI-Driven Drug Discovery - NVIDIA Newsroom	2026-01-12	113	2026-01-12	183.22000122070312	187.11999511718750	183.02000427246094	184.94000244140625	13796850.0	True
12	NVIDIA Blog	NVIDIA Brings GeForce RTX Gaming to More Devices With New GeForce NOW Apps for Linux PC and Amazon Fire TV - NVIDIA Blog	2026-01-05	120	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
13	NVIDIA Blog	NVIDIA DGX SuperPOD Sets the Stage for Rubin-Based Systems - NVIDIA Blog	2026-01-05	72	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
14	NVIDIA Newsroom	NVIDIA Announces Alpamayo Family of Open-Source AI Models and Tools to Accelerate Safe, Reasoning-Based Autonomous Vehicle Development - NVIDIA Newsroom	2026-01-05	152	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
15	NVIDIA Blog	Steel, Sensors and Silicon: How Caterpillar Is Bringing Edge AI to the Jobsite - NVIDIA Blog	2026-01-07	92	2026-01-07	188.57000732421875	191.36999511718750	186.55999755859372	189.11000061035156	15354320.0	True
16	NVIDIA Blog	CEOs of NVIDIA and Lilly Share 'Blueprint for What Is Possible' in AI and Drug Discovery - NVIDIA Blog	2026-01-13	102	2026-01-13	185.0	188.11000061035156	183.39999389648438	185.80999755859372	16012890.0	True
17	NVIDIA Blog	NVIDIA RTX Accelerates 4K AI Video Generation on PC With LTX-2 and ComfyUI Upgrades - NVIDIA Blog	2026-01-05	97	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
18	NVIDIA Newsroom	NVIDIA BlueField-4 Powers New Class of AI-Native Storage Infrastructure for the Next Frontier of AI - NVIDIA Newsroom	2026-01-05	117	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
19	NVIDIA Blog	NVIDIA DLSS 4.5, Path Tracing and G-SYNC Pulsar Supercharge Gameplay With Enhanced Performance and Visuals - NVIDIA Blog	2026-01-05	120	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
20	NVIDIA Blog	GeForce NOW Rings In 2026 With 14 New Games in January - NVIDIA Blog	2026-01-01	68							False
21	NVIDIA Blog	NVIDIA Expands Global DRIVE Hyperion Ecosystem to Accelerate the Road to Full Autonomy - NVIDIA Blog	2026-01-05	100	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
22	NVIDIA Blog	NVIDIA DRIVE AV Software Debuts in All-New Mercedes-Benz CLA - NVIDIA Blog	2026-01-05	74	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
23	NVIDIA Developer	Introducing NVIDIA BlueField-4-Powered Inference Context Memory Storage Platform for the Next Frontier of AI - NVIDIA Developer	2026-01-06	127	2026-01-06	190.52000427246094	192.16999816894531	186.82000732421875	187.24000549316406	17686260.0	True
24	Business Insider	I quit Nvidia before my stock blew up, but Jensen Huang taught me career lessons I still follow today - Business Insider	2026-01-20	120	2026-01-20	181.89999389648438	182.38000488281250	177.61000061035156	178.07000732421875	22334530.0	True
25	NVIDIA	Out Now: NVIDIA RTX Remix Update Introduces New Logic System For Dynamic Graphics - NVIDIA	2026-01-28	90							False
26	NVIDIA Developer	NVIDIA DLSS 4.5 Delivers Super Resolution Upgrades and New Dynamic Multi Frame Generation - NVIDIA Developer	2026-01-14	108	2026-01-14	184.32000732421875	184.46000671386720	180.80000305175780	183.13999938964844	15958610.0	True
27	qz.com	Nvidia and Mercedes made an AV. I went along for the ride - qz.com	2026-01-06	66	2026-01-06	190.52000427246094	192.16999816894531	186.82000732421875	187.24000549316406	17686260.0	True
28	NVIDIA Blog	NVIDIA Unveils New Open Models, Data and Tools to Advance AI Across Every Industry - NVIDIA Blog	2026-01-05	96	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
29	NVIDIA	GeForce Game Ready Driver For The ARC Raiders: Headwinds Update - NVIDIA	2026-01-27	72	2026-01-27	187.24000549316406	190.0	185.69999694824220	188.52000427246094	14323350.0	True
30	NVIDIA Blog	'Largest Infrastructure Buildout in Human History': Jensen Huang on AI's 'Five-Layer Cake' at Davos - NVIDIA Blog	2026-01-21	113	2026-01-21	179.05003051757813	185.38000488281250	178.39999389648438	183.32000732421875	20038100.0	True
31	Commonwealth Fusion Systems	Commonwealth Fusion Systems Accelerates Commercial Fusion With Siemens and NVIDIA, Leveraging AI-Powered Digital Twins - Commonwealth Fusion Systems	2026-01-06	148	2026-01-06	190.52000427246094	192.16999816894531	186.82000732421875	187.24000549316406	17686260.0	True
32	Microsoft Azure	Microsoft's strategic AI datacenter planning enables seamless, large-scale NVIDIA Rubin deployments - Microsoft Azure	2026-01-05	117	2026-01-05	191.75999450683597	193.63000488281250	186.1499938964844	188.1199951171875	18352970.0	True
33	NVIDIA Blog	How to Get Started With Visual Generative AI on NVIDIA RTX PCs - NVIDIA Blog	2026-01-22	76	2026-01-22	184.75	186.16999816894531	183.52999267578125	184.83999633789065	13936600.0	True
34	Eli Lilly	NVIDIA and Lilly Announce Co-Innovation AI Lab to Reinvent Drug Discovery In the Age of AI - Eli Lilly	2026-01-12	102	2026-01-12	183.22000122070312	187.11999511718750	183.02000427246094	184.94000244140625	13796850.0	True

4.3 Aggregating news and computing sentiment features

For modelling purposes, I want to work at a daily level rather than at the individual headline level. To do this, I construct a daily_news DataFrame by grouping the news data by pubdate and computing aggregate statistics. In particular, I use groupby("pubdate").agg({"sentiment_score": ["mean", "count"]}) to calculate the mean sentiment score and the number of headlines per day. The resulting column names are then flattened to Date, sentiment_mean, and news_count.

The sentiment_score field is intended to quantify the polarity of each headline, with positive values representing more positive language and negative values representing more negative language. Although the exact sentiment computation code is not fully shown in the snippet, the design of the pipeline assumes that each headline obtains a numeric sentiment score from an NLP tool, which is then averaged across all headlines for a given day.
Together, sentiment_mean and news_count provide a compact summary of the news environment for each trading day.

4.4 Building the final modelling dataset

To link the daily news features with the stock market data, I merge the daily_news DataFrame with the stock DataFrame on the Date column using an inner join. After merging, I sort rows by date to preserve time order and create a new column next_close by shifting the Close price one day ahead. This allows me to compare the current closing price with the next day's closing price.

The binary target variable target is then defined as 1 if next_close is greater than the current Close, and 0 otherwise. This turns the problem into a daily classification task:

predict whether the next day's closing price will go up relative to today. Any rows with missing values after shifting are dropped, and the resulting final dataset is saved as final_model_data.csv. The feature matrix X consists of sentiment_mean, news_count, Open, Close, and Volume, while the label vector y is the target column.

---

5. Modelling Approach

5.1 Time-series cross-validation

Financial time series have a natural chronological order, and using standard random cross-validation can introduce look-ahead bias. To address this, I use TimeSeriesSplit from scikit-learn to create ordered training and test folds. The number of splits is set dynamically as n_splits = max(2, min(5, len(final_df) // 10)), ensuring that there are enough observations per fold while keeping between two and five splits.

In each fold, earlier dates are used for training and later dates for testing, preserving the direction of time. This setup better reflects a realistic scenario where models are fit on past data and evaluated on future unseen data. It also reduces the risk of overly optimistic performance estimates that can occur when information from the future leaks into the training set.

5.2 Models and features

I compare two tree-based classification models that are commonly used in applied machine learning: Random Forest and XGBoost. For both models, I use the same set of input features: daily average sentiment (sentiment_mean), number of headlines (news_count), and the stock price-related features Open, Close, and Volume. These features combine information from the news flow with basic market data.

The RandomForestClassifier is configured with 100 trees, a maximum depth of 5, and class_weight='balanced' to account for any imbalance between up and down days. The XGBClassifier uses 100 estimators, a maximum depth of 3, a learning rate of 0.05, and eval_metric='logloss'. Both models are initialised with a fixed random seed to make the results reproducible. For each fold, I fit each model on the training portion and then obtain both class predictions and predicted probabilities on the test portion.

	Date	sentiment_mean	news_count	Close	High	Low	Open	Volume	next_close	target
2	2024-01-30	0.5670557022094727	1	62.73904800415039	63.457649460395544	62.225334241470755	62.86498111554158	410735000	61.49274444580078	0
3	2024-01-31	0.0	1	61.49274444580078	62.23433190687522	60.666205224632265	61.40579108468123	453795000	69.29339599609375	1
4	2024-02-05	0.0	1	69.29339599609375	69.45830503749701	67.16758142852466	68.18701012748092	680078000	68.18501281738281	0
5	2024-02-06	0.0	1	68.18501281738281	69.71515858330277	66.26308702618283	69.59122707783612	683111000	69.60221099853516	1
6	2024-02-08	0.0	1	69.60221099853516	70.75456869186667	69.41631758373076	70.03496848915448	414422000	72.20777130126953	1
7	2024-02-12	0.0	2	72.20777130126953	74.56945447682682	71.21032591879988	72.55957267433736	613710000	72.08784484863281	0
8	2024-02-13	0.4070650637149811	6	72.08784484863281	73.40910749101484	69.58124543898839	70.36080999462226	602580000	73.8588638305664	1
9	2024-02-14	0.13630921840667726	5	73.8588638305664	74.19467532641171	71.89795802434332	73.16125433620005	504917000	72.61753845214844	0
10	2024-02-15	0.030662208795547485	8	72.61753845214844	73.93380643951798	72.35968650824275	73.82787028335872	420122000	72.57256317138672	0
11	2024-02-16	0.0	2	72.57256317138672	74.36056840015162	72.46062602851259	74.05873462198204	495327000	69.41332244873047	0
12	2024-02-20	0.26930707693099976	2	69.41332244873047	71.91592565937303	67.696276964279	71.90692798783778	704833000	67.43444061279297	0
13	2024-02-21	0.42102062351563396	17	67.43444061279297	68.84965273678795	66.21112295055113	67.9681399359681	673755000	78.49427032470703	1
14	2024-02-22	0.2029419869184494	20	78.49427032470703	78.53124465919488	74.17867388920303	74.9832259524774	865100000	78.77310943603516	1
15	2024-02-23	0.12876980006694794	8	78.77310943603516	82.34811331480627	77.52680222303147	80.74501027740159	829388000	79.04796600341797	1
16	2024-02-26	0.25110953194754465	7	79.04796600341797	80.60110162092384	78.46128722131758	79.65562156391138	503973000	78.65717315673828	0
17	2024-02-27	0.14010360836982727	6	78.65717315673828	79.43574633865495	77.11903628247967	79.3367948159067	391705000	77.61975860595703	0
18	2024-02-28	0.0	7	77.61975860595703	78.88904810428491	77.08205615581203	77.57678328073024	393110000	88.65476989746094	1
19	2024-03-06	0.0	1	88.65476989746094	89.67824984643924	86.9856232402923	87.97712182383515	582520000	87.4833755493164	0
20	2024-03-08	0.0	2	87.4833755493164	97.35034416273697	86.46189342253078	95.08949635927887	1142269000	85.73026275634766	0
21	2024-03-11	0.08003643155097961	6	85.73026275634766	88.75171621511622	84.1230810265849	86.38492752767289	678364000	91.86614227294922	1
22	2024-03-12	0.0	1	91.86614227294922	91.91311538288066	86.10607991861596	88.00411383105026	668075000	90.8416519165039	0
23	2024-03-13	0.0	1	90.8416519165039	91.45733510152752	88.38989990202793	91.00856657140831	635713000	87.89915466308594	0

5.3 Evaluation metrics

To evaluate performance, I compute a set of standard classification metrics. Accuracy measures the overall fraction of correctly predicted days, while precision measures the fraction of predicted “up” days that actually went up. Recall measures the fraction of actual “up” days that were correctly detected, and the F1-score provides a harmonic mean of precision and recall, which is useful when the classes are imbalanced.

In addition, I calculate the ROC-AUC score based on the predicted probabilities, which summarises the trade-off between true positive rate and false positive rate across different decision thresholds. These metrics are computed for each model and fold using scikit-learn functions such as accuracy_score, precision_score, recall_score, f1_score, and roc_auc_score. The per-fold results are stored in a DataFrame and written to experiment_results_comparison.csv, and an aggregated summary by model is saved in final_model_summary.csv.

	fold	model	accuracy	precision	recall	f1	roc_auc
2	1	RandomForest	0.5416666666666666	0.53125	0.918918918918919	0.6732673267326733	0.5544401544401545
3	1	XGBoost	0.5416666666666666	0.53125	0.918918918918919	0.6732673267326733	0.5575289575289575
4	2	RandomForest	0.5555555555555556	0.5625	0.5	0.5294117647058824	0.5478395061728396
5	2	XGBoost	0.5277777777777778	0.5178571428571429	0.8055555555555556	0.6304347826086957	0.6064814814814815
6	3	RandomForest	0.5555555555555556	0.5625	0.7105263157894737	0.627906976744186	0.601780185758514
7	3	XGBoost	0.5277777777777778	0.5454545454545454	0.631578947368421	0.5853658536585366	0.5917182662538699
8	4	RandomForest	0.3888888888888889	0.42857142857142855	0.06976744186046512	0.12	0.4274258219727346
9	4	XGBoost	0.4166666666666667	1.0	0.023255813953488372	0.045454545454545456	0.42662389735364875
10	5	RandomForest	0.5277777777777778	0.6	0.08571428571428572	0.15	0.44942084942084937
11	5	XGBoost	0.4861111111111111	0.42857142857142855	0.17142857142857143	0.24489795918367346	0.48764478764478764

---

6. Results and Discussion

The experiment_results_comparison.csv file contains multiple rows per model, one for each time-series fold, with columns for accuracy, precision, recall, F1, and ROC-AUC. Aggregating these results by model in final_model_summary.csv gives an average view of performance over time. Although exact numeric values will depend on the specific dataset realisation, the summary allows me to compare Random Forest and XGBoost on a like-for-like basis.

In general, I would expect the two models to achieve accuracies modestly above 0.5 if they are capturing some signal beyond random guessing, although financial direction prediction is notoriously challenging. By inspecting the summary, I can identify which model has higher accuracy and F1-score and whether one model tends to have better recall (catching more up moves) at the cost of precision, or vice versa. ROC-AUC values close to 0.5 would indicate near-random performance, while higher values suggest that the predicted probabilities have useful ranking ability.

Beyond raw scores, it is important to interpret how the features might be contributing to the predictions. The inclusion of Open, Close, and Volume means that each model can learn patterns from recent price levels and trading activity. The sentiment features sentiment_mean and news_count provide an additional channel, capturing the tone and intensity of the news flow, which may help improve predictions on days with particularly positive or negative news. However, given the limited scope of the data and the simplicity of the sentiment signal, there is also a risk that the models rely more heavily on price-based features than on news sentiment.

	model	accuracy	precision	recall	f1	roc_auc
2	RandomForest	0.513888888888889	0.5369642857142857	0.45698539245662867	0.4201172136365483	0.5161813035530184
3	XGBoost	0.5	0.6046266233766233	0.5101475614449911	0.4358840935276249	0.533999478052549

7. Limitations and Future Work

This project has several limitations that should be kept in mind when interpreting the results. First, it focuses on a single stock, Nvidia, over a relatively short period of around two years, which restricts the amount of training data and the diversity of market conditions included. Second, the news data consists only of headlines, not full article texts, so the sentiment scores are based on very short pieces of text and may be noisy or incomplete representations of the underlying news.

Third, the prediction target is a simple binary indicator of whether the next day's closing price is higher than the current close. This ignores the magnitude of the move, transaction costs, and practical trading constraints, so even a model with modest predictive power might not translate into a profitable strategy. Additionally, the models do not incorporate other potentially relevant information such as technical indicators, market indices, macroeconomic variables, or alternative text sources like social media.

There are several ways in which the work could be extended. On the data side, one could collect longer histories and include multiple stocks or indices to test whether the approach generalises beyond Nvidia. On the NLP side, more advanced methods such as transformer-based language models could be used to derive richer sentiment or topic features from full news articles rather than just headlines. On the modelling side, one could experiment with sequence models, additional regularisation, or ensemble techniques, and also perform more thorough hyperparameter tuning.

---

8. Conclusion

In this project, I constructed an end-to-end pipeline that links Nvidia-related news headlines with Nvidia's daily stock prices and uses this combined dataset to predict next-day price direction. The pipeline begins by scraping approximately two years of Nvidia headlines from Google News, cleaning and structuring them, and downloading matching NVDA price data from Yahoo Finance. I engineered daily features that summarise the news flow and merged them with stock prices to form a supervised learning dataset.

Using this dataset, I trained and evaluated Random Forest and XGBoost classifiers under a time-series cross-validation scheme, comparing them on metrics such as accuracy, F1-score, and ROC-AUC. The exercise demonstrates the feasibility of integrating textual sentiment with market data and illustrates some of the challenges of short-term stock movement prediction. While the models provide some insight into the relationship between news and price changes, the results also emphasise the

need for richer data, more advanced NLP methods, and careful evaluation before such models could be used in a real trading context.