

News-Driven Quantitative Trading System Using Sentiment Analysis

A Comprehensive Quantitative Finance Project Integrating NLP, Market Data, and Machine Learning

1. Introduction

Financial markets are fundamentally driven by information. Every earnings announcement, analyst upgrade, macroeconomic release, regulatory update, or breaking news headline has the potential to shift investor expectations and, consequently, asset prices. In modern electronic markets, this information is disseminated at a speed and scale far beyond direct human processing capability, creating a strong incentive for **systematic, data-driven trading strategies** that can ingest, interpret, and react to new information in a consistent and repeatable manner.

Over the last two decades, advances in computing power, machine learning, and natural language processing (NLP) have transformed how market participants analyze information. Textual data—once considered qualitative, subjective, and difficult to quantify—can now be systematically structured and incorporated into predictive models. This project is situated at the intersection of these technological and methodological developments.

This work presents a **news-driven quantitative research and modeling pipeline** that integrates **NLP-based sentiment analysis** with **statistical and machine-learning models** to predict short-term stock price movements. The project focuses on NVIDIA Corporation (NVDA), a highly liquid, large-capitalization equity that is particularly sensitive to both technological developments and macro-economic narratives, making it well suited for sentiment-based analysis.

The system is intentionally designed to resemble a real-world quantitative research workflow rather than a simplified academic exercise:

- Raw news ingestion and preprocessing
- Sentiment extraction with confidence weighting
- Alignment of unstructured news data with structured historical price data
- Construction of forward-looking, leakage-free prediction targets
- Model training under realistic temporal constraints
- Rigorous evaluation using multiple performance metrics
- Explicit comparison of model predictions with realized market outcomes

Rather than optimizing purely for maximum short-term profitability, the primary objective of this project is **methodological correctness, interpretability, and robustness**. In professional quantitative finance, a modest but stable signal is often far more valuable than an unstable, overfit model that performs well only in hindsight.

2. Data Sources and Collection

2.1 News Data

The news dataset consists of structured news articles related to NVIDIA. Each record contains the following key fields:

- **Source:** The originating publisher of the news article
- **Headline:** A concise textual summary capturing the essence of the event
- **Publication Date (pubdate):** The timestamp indicating when the news became publicly available

Headlines are particularly useful in quantitative research because they compress the core informational content of an article into a short, signal-dense format. This makes them computationally efficient to process while retaining much of the information that drives short-term market reactions.

Moreover, headlines often reflect how information is framed. In financial markets, framing effects—such as emphasis on risks versus opportunities—can be as influential as the underlying facts themselves.

2.2 Market Data

Market data is collected using the `yfinance` library and consists of daily OHLCV (Open, High, Low, Close, Volume) price data for NVDA. This dataset serves as the *ground truth* against which all model predictions are evaluated.

Daily frequency data provides a practical balance between signal and noise. While intraday data offers greater granularity, it introduces significant microstructure noise and execution complexity. Lower-frequency data, on the other hand, may dilute the immediate impact of news. For this project, daily data represents an appropriate and realistic compromise.

2.3 Temporal Alignment and Data Integrity

A central challenge addressed in this project is the **temporal alignment** of news and market data. All explanatory features are constructed using only information available at time t , while prediction targets are defined using price movements at time $t+1$.

This strict temporal separation is essential. It eliminates look-ahead bias and ensures that the reported model performance reflects what could realistically have been achieved in a live forecasting or trading environment.

3. Feature Engineering

3.1 Headline-Based Sentiment Features

Two primary features are derived from each news headline:

1. Sentiment Direction

Each headline is classified as *positive*, *neutral*, or *negative* using a finance-aware sentiment analysis

model. Unlike general-purpose sentiment tools, financial sentiment models are trained to recognize domain-specific language such as earnings surprises, guidance revisions, and risk disclosures.

2. Sentiment Confidence

Alongside the sentiment label, the model produces a confidence score that measures how strongly the sentiment is expressed. This allows the system to differentiate between weakly worded opinions and strongly expressed views.

These components are combined into a single **continuous sentiment score**:

$$\text{Sentiment Score} = \text{Sentiment Direction} \times \text{Sentiment Confidence}$$

This formulation preserves both direction and intensity, which is particularly important in financial contexts where not all positive or negative news carries equal informational weight.

3.2 Market-Based Features

In addition to sentiment, several market features are incorporated: - Open price - Close price - Trading volume

These variables provide context regarding prevailing price levels, recent market behavior, and liquidity conditions. In practice, the market's reaction to news often depends on existing trends and participation levels.

3.3 Feature Design Philosophy

The feature set is intentionally compact and interpretable. In quantitative finance, indiscriminate feature expansion frequently leads to overfitting and weak out-of-sample performance. This project prioritizes **economic intuition, signal clarity, and interpretability** over sheer feature count.

4. Target Construction

4.1 Forward-Looking Returns

The prediction target is based on **next-day returns**, defined as:

$$\text{Return}_{t+1} = \frac{\text{Close}_{t+1} - \text{Close}_t}{\text{Close}_t}$$

This formulation ensures that model predictions are evaluated strictly against future, realized outcomes rather than contemporaneous price movements.

4.2 Binary Classification Target

To reduce noise and reflect realistic trading constraints, the continuous return is converted into a binary classification target:

- **1 (Positive Move):** Next-day return greater than 0.3%
- **0 (Non-Positive Move):** Otherwise

The threshold removes minor price fluctuations that are unlikely to be economically meaningful after accounting for transaction costs and slippage.

4.3 Class Imbalance Considerations

As is common in financial return series, the resulting target distribution is imbalanced, with a greater number of non-positive outcomes. This imbalance is explicitly handled during model training through class-weighted learning, ensuring that positive outcomes receive appropriate emphasis.

5. Modeling Approach

5.1 Problem Formulation

The task is framed as a **binary classification problem**:

Given news sentiment and market context at time t , predict whether the stock price will experience a meaningful positive move at time $t+1$.

This framing closely mirrors real-world trading decisions, such as whether to initiate a long position or remain neutral.

5.2 Model Selection and Rationale

A **Random Forest classifier** is selected as the primary model due to its: - Ability to capture nonlinear interactions between sentiment and price behavior - Robustness to noisy inputs - Strong empirical performance on tabular financial datasets - Lower overfitting risk compared to deep learning models in small-sample settings

The model is configured with controlled tree depth, a sufficiently large ensemble size, and balanced class weights, reflecting best practices in exploratory quantitative research.

6. Validation Methodology

6.1 Cross-Validation Strategy

The project employs **Stratified K-Fold cross-validation** to ensure that:

- Each fold preserves the original class distribution
- Performance estimates are not dominated by a single subsample

Although time-series cross-validation is often preferred in production trading systems, stratified folds are appropriate here given the dataset size and the focus on classification behavior rather than execution timing.

6.2 Evaluation Metrics

Model performance is assessed using multiple complementary metrics:

- **Accuracy:** Overall correctness
- **Precision:** Reliability of positive predictions
- **Recall:** Ability to capture actual positive moves
- **F1-Score:** Balance between precision and recall
- **ROC-AUC:** Quality of probabilistic ranking

Using a multi-metric evaluation framework avoids misleading conclusions that could arise from accuracy alone, particularly in imbalanced datasets.

7. Prediction vs. Actual Outcome Analysis

A defining strength of this project is the explicit comparison of **model predictions with realized market outcomes**. For each test observation, the system records:

- Predicted probability
- Predicted class label
- Actual realized target
- Correctness indicator

This produces a transparent audit trail that enables:

- Detailed inspection of model errors
- Identification of market regimes where performance deteriorates
- Post-hoc analysis suitable for strategy simulation and risk assessment

Such traceability is a critical requirement in professional quantitative research and is often missing from purely academic implementations.

8. Results and Discussion

Empirical results indicate that the model achieves:

- Predictive performance above random guessing
- Non-zero precision and recall, confirming the presence of exploitable signal
- Reasonably stable behavior across validation folds

These results highlight the intrinsic difficulty of short-horizon price prediction. In financial markets, even small improvements over chance can be economically meaningful when applied consistently and with appropriate risk controls.

Crucially, the project prioritizes **statistical validity and robustness** over headline-grabbing performance metrics, aligning closely with industry-grade quantitative research standards.

9. Limitations and Future Extensions

Several limitations are acknowledged:

- The single-asset focus limits generalization
- Daily frequency may miss intraday sentiment effects
- The feature set is intentionally minimal

Potential future extensions include:

- Expansion to a multi-asset universe
- Intraday or event-level sentiment modeling
- Full trading strategy backtesting with transaction costs
- Integration of macroeconomic indicators and derivatives-based signals

10. Conclusion

This project delivers a complete, end-to-end **news-driven quantitative research pipeline**. By combining NLP-based sentiment analysis with disciplined statistical modeling and rigorous validation, it demonstrates how unstructured textual information can be transformed into actionable financial signals.

More importantly, the project emphasizes methodological rigor, transparency, and realism—qualities that define professional quantitative finance research. While not intended as a production-ready trading system, it provides a strong and extensible foundation upon which more advanced models and strategies can be developed.