Can Analysts Predict Markets? Using Equity Research Reports to Forecast Gilead Sciences Stock Price

UC Berkeley College of Computing, Data Science, and Society

Qamil Mirza[1], Tara Timm[1], Holden Carrillo[1], Zhen Liu[1], Amy Tran[1], Ethan Yen[2], Eric Mecca[2]

[1]University of California, Berkeley [2]Gilead Sciences

Overview

Forecasting stock movements is difficult because markets quickly incorporate new information, leaving limited signal in historical prices alone. For biotechnology firms like Gilead Sciences (GILD), analyst reports may contain forward-looking insights about clinical developments, earnings expectations, and investor sentiment. This project extracts sentiment from sell-side research reports, aligns it with daily GILD market data, and tests whether it improves walk-forward forecasting performance over market-only baselines. Ultimately, we evaluate whether analyst sentiment anticipates future price movements or simply reflects information already priced into the market.

Building Sentiment Features

In Fall 2025, Gilead's Data Discovery team built a pipeline to measure sentiment in analyst reports using LLMs (Claude, Llama, and FinBERT) and tested how different design choices, such as chunk size, prompting, and model selection, affect performance. A large majority of reports were positive, with almost every firm averaging above a neutral rating of 4.

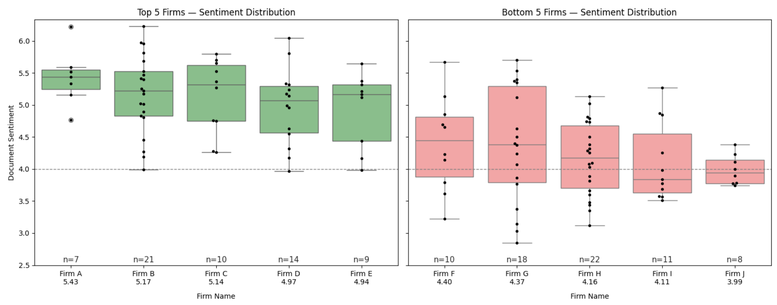

Figure 1. 5 Firms With Highest and Lowest Average Sentiment Towards Gilead, 2024-25

Finding a Baseline

Before determining analyst impact, we created a baseline model to predict prices

- **Naive:** Use yesterday's close price to predict the next day's price
- **Mean:** Use the historical average of past prices to predict the next day's price
- **Linear:** Fit a linear model to past prices to predict the next day's price
- **ARIMA:** Fit an ARIMA model to past prices; the most optimal model ended up being ARIMA(0,1,0) over most windows, which is the naive model

The Naive model outperformed all initial baselines.

Model	Mean	Median	STD
Naive	3.543	3.446	2.017
Mean	5.381	4.947	3.176
Linear	4.010	3.361	2.869
ARIMA	4.581	4.096	2.718

Table 1. Performance of various baseline models we implemented, all metrics are of RMSE over all training windows

The XGBoost Model

Next, we introduce XGBoost, a widely used ensemble model used with structured financial data. XGBoost constructs a sequence of decision trees, where each successive tree is trained to correct the errors of the previous one. By applying gradient boosting to this framework, the model captures nonlinear relationships and complex feature interactions that are often present in market data.

For our baseline model, we trained the XGBoost model with a combination of numerical and categorical market features, such as lagged returns, moving averages, and days of the week. Specifically, we establish a trading window and train our model using 120 days of past data. Then, we generate price predictions for 10 trading days into the future, and repeat this process for our entire time-series dataset.

Model performance was evaluated using the Root Mean Squared Error (RMSE), an out-of-sample performance metric commonly used for evaluating systematic trading models. Improvements over this baseline would suggest that analyst sentiment contains meaningful signals that are not captured by historical price data.

Adding Sentiment

With our baseline benchmark established, we train new XGBoost models with baseline features and the sentiment scores of one firm. We generate price predictions for 10 trading days, and the process is repeated for all firms in our dataset.

Performance against our baseline is evaluated based on % improvements in mean RMSE over our baseline model and cross-validated by visualizing the price predictions of each firm-augmented XGBoost model across all forecasting windows.

Figure 2. A forecasting window, with blue representing training data, dotted lines representing firm-augmented model predictions, and black representing the ground-truth close price of GILD.

Results

In total, 20 out of the 34 firms evaluated provided an improvement over the baseline, with the largest improvement being 0.23%.

XGB RMSE	Improvement vs Baseline	% Improvement vs Baseline	Beats Baseline	% Improvement vs Naive (context)
3.1343	+0.0071	+0.23%	Yes	+4.26%
3.1350	+0.0064	+0.20%	Yes	+4.24%
3.1365	+0.0049	+0.16%	Yes	+4.19%
3.1368	+0.0046	+0.15%	Yes	+4.18%
3.1368	+0.0046	+0.15%	Yes	+4.18%
3.1517	-0.0103	-0.33%	No	+3.73%
3.1501	-0.0087	-0.28%	No	+3.78%
3.1499	-0.0085	-0.27%	No	+3.78%
3.1479	-0.0065	-0.21%	No	+3.84%
3.1462	-0.0048	-0.15%	No	+3.89%

Figure 3. Baseline Reference & Top 5 / Bottom 5 Performances

Significance Testing

Diebold–Mariano Test

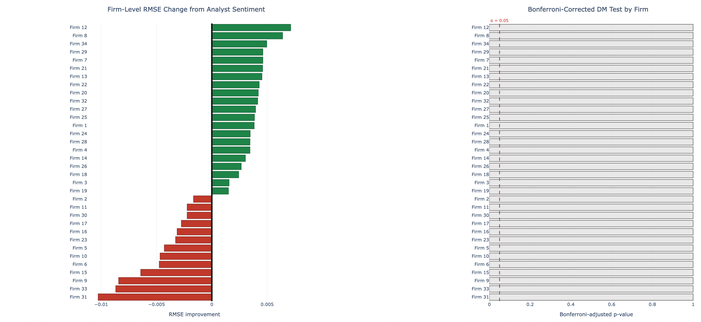

Figure 4. Left: Firm-level RMSE improvement from sentiment. Right: Bonferroni-corrected Diebold–Mariano p-values. Several firms reduce RMSE, but none remain significant after multiple-comparison correction.

Noise Permutation Test

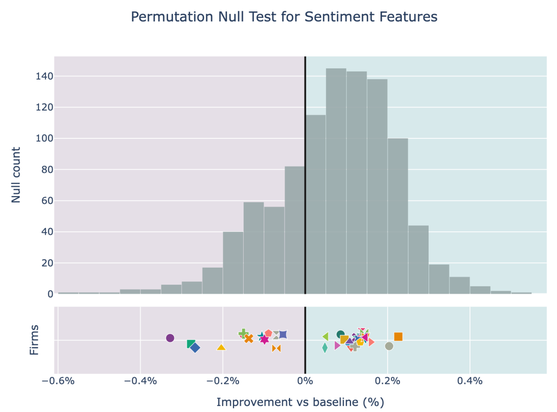

Figure 5. Noise-feature permutation test comparing sentiment gains against a random-feature null. No firm significantly outperforms the null at $\alpha = 0.05$.

Conclusions

Analyst sentiment did not show statistically significant predictive value for GILD price forecasting beyond market-only baselines. Still, this work develops a systematic framework for evaluating analyst reports as quantitative signals, allowing comparison across firms through walk-forward ablations and statistical testing. Future work should explore recency-weighted sentiment, event-specific report filtering, firm-level credibility weights, and alternative targets such as volatility, abnormal returns, or post-report price reactions.

Acknowledgments

We thank Ethan Yen and Eric Mecca for their guidance and mentorship throughout this project, and we gratefully thank the Gilead Sciences leadership team for their support and feedback.