Evaluating Large Language Models for Financial News Sentiment under Market Frictions

The 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)

Kemal Kirtac
Department of Computer Science
University College London
66–72 Gower Street, London WC1E 6EA, United Kingdom
kemal.kirtac.21@ucl.ac.uk
https://www.ucl.ac.uk/computer-science

Abstract

This paper studies whether large language models can extract useful sentiment signals from firm-specific financial news when evaluation accounts for realistic market frictions. Many financial NLP studies report strong offline prediction results, but these do not always show whether model outputs remain useful once trading constraints are imposed. I address this gap by evaluating sentiment models through classification performance, return predictability, and implementable portfolio performance. The analysis links Refinitiv News Analytics to CRSP and begins with 3,129,924 U.S. news items published between January 1, 2010 and January 30, 2026. Filtering retains single-firm stories, removes redundant coverage using a five-day cosine-similarity novelty screen, and restricts the sample to tradable stocks with positive bid and ask quotes, minimum share and dollar volume thresholds, quoted spreads below 20%, and available Amihud illiquidity ratios and Kyle's lambda estimates. The final sample contains 973,481 tradable news items linked to 3,452 firms. I compare six sentiment approaches: LLaMA–3, OPT, RoBERTa, BERT, FinBERT, and the Loughran–McDonald dictionary. LLaMA–3 achieves the strongest classification performance with 78.2% accuracy and produces the largest predictive coefficients in panel regressions. Daily rebalanced long–short portfolios with a 5 bps trading cost show that the LLaMA–3 strategy earns a cumulative return of approximately 180% from June 2024 to January 2026, followed by OPT with 155% and RoBERTa with 120%, while the dictionary-based strategy loses 9%. The results show that evaluation becomes more informative when financial NLP models are assessed beyond offline accuracy and under realistic deployment constraints. High-capacity language models retain economically meaningful predictive content under market frictions, whereas simpler lexicon-based methods do not.

1 Introduction

Textual information plays an increasingly central role in empirical finance. Prior work shows that news coverage, corporate disclosures, and investor communication contain predictive information about asset prices and firm performance (Tetlock, 2007a; Tetlock et al., 2008; Price et al., 2015; Huang et al., 2014; Li, 2008). Studies of regulatory filings, earnings calls, and online platforms further demonstrate that sentiment and attention extracted from text shape trading behavior and return dynamics (Loughran and McDonald, 2011; Da et al., 2011; Chen et al., 2014). This literature establishes that financial text contains economically meaningful signals relevant for asset pricing.

Most empirical applications still operationalize text using relatively simple sentiment measures, most commonly dictionary-based word counts (Loughran and Mcdonald, 2011; Malo et al., 2014). These approaches ignore context, syntax, and semantic structure, which limits their ability to capture nuanced information embedded in firm-specific news. Existing studies also rarely evaluate whether text-based signals remain economically valuable once realistic trading frictions are imposed. Liquidity constraints, bid–ask spreads, price impact, and limits to trading capacity can prevent investors from exploiting information immediately, weakening the link between textual signals and realized returns (Kyle, 1985; Amihud, 2002; Pastor and Stambaugh, 2003).

This paper studies LLM-based news sentiment in portfolio construction within a framework that embeds market frictions directly into sample construction and strategy simulation. The analysis links Refinitiv news to CRSP returns and begins with 3,129,924 U.S. news items published between January 1, 2010 and January 30, 2026. Filtering retains 1,985,135 single-firm stories, removes redundant coverage using a five-day cosine-similarity nov-

Proceedings of the Workshop on Evaluating Evaluations (EvalEval), pages 26–35

elty screen, and produces 1,122,475 unique news items. Liquidity and microstructure screens then restrict the universe to stocks with strictly positive bid and ask quotes, minimum share and dollar volume thresholds, quoted spreads below 20%, and the availability of Amihud illiquidity ratios and Kyle's lambda price-impact estimates. The final sample contains 973,481 tradable news items linked to 3,452 firms, which ensures that sentiment signals are evaluated only where institutional execution is feasible.

The empirical strategy proceeds in three steps. First, each model's sentiment score is evaluated for its ability to classify the sign of the subsequent three-day excess return, providing a direct measure of return-relevant information in news text. Second, predictive regressions with firm and date fixed effects test whether sentiment scores forecast next-day returns under two-way clustered inference. Third, daily rebalanced value-weighted long, short, and long–short portfolios translate cross-sectional sentiment rankings into positions while incorporating transaction costs, timestamp-based execution aligned with news arrival, participation limits tied to daily dollar volume, and microstructure-informed tradability screens. I compare the sentiment-based strategies with buy-and-hold benchmarks based on the Dow Jones Industrial Average and the Nasdaq Composite.

The results show large and systematic differences across sentiment extraction methods. Transformer-based LLM sentiment measures generate stronger predictability than the dictionary benchmark. LLaMA–3 delivers the highest classification accuracy and the strongest predictive coefficients, followed by OPT and RoBERTa, while the Loughran–McDonald dictionary performs close to chance. Trading simulations show that these informational differences translate into economically meaningful performance under realistic frictions, with LLaMA–3-based long–short portfolios delivering the strongest cumulative gains over the out-of-sample evaluation window.

This paper contributes to evaluation research by showing that financial NLP systems should be assessed not only by offline classification accuracy, but also by downstream return predictability and implementable portfolio performance under market frictions. It also contributes to the finance literature that integrates machine learning and text analysis into asset-pricing applications (Jegadeesh and Wu, 2013; Manela and Moreira, 2017; Ke et al., 2020).

Sample construction step	Count
All Refinitiv news items	3,129,924
Single-firm news items	1,985,135
After 5-day novelty filter	1,122,475
After liquidity and microstructure filters	973,481
Unique firms in final sample	3,452

Table 1: Sample construction and filtering steps for the Refinitiv–CRSP merged dataset.

The core contribution lies in showing how LLM-derived news sentiment can be operationalized for portfolio construction and evaluated under realistic execution constraints, while highlighting how model architecture and context sensitivity affect the economic value of textual signals.

The remainder of the paper proceeds as follows. Section 2 describes the data and sample construction. Section 3 outlines the sentiment extraction and empirical methodology. Section 4 presents the results. Section 5 concludes.

2 Data

The analysis combines firm-level stock data from the Center for Research in Security Prices (CRSP) with firm-specific news from the Refinitiv News Analytics archive. CRSP provides daily returns, prices, trading volume, shares outstanding, and market capitalization for U.S. equities listed on the NYSE, NASDAQ, and AMEX. Refinitiv supplies time-stamped news articles and alerts linked to publicly traded firms. Merging these sources allows sentiment extracted from news text to be mapped directly to subsequent stock returns.

The initial sample includes all Refinitiv articles linked to at least one U.S. equity between January 1, 2010 and January 30, 2026, yielding 3,129,924 news items associated with U.S. listed firms. The sample retains only articles linked to a single firm to ensure unambiguous return attribution and requires a valid three-day excess return computed from CRSP data. A novelty screen based on cosine similarity removes redundant coverage by excluding any article that exceeds 0.80 similarity to an earlier story published within the previous five trading days. This procedure produces 1,122,475 unique news items.

Liquidity and market-friction filters further restrict the sample to stocks that plausibly support institutional trading and allow microstructure variables to be computed reliably. The filters require

strictly positive bid and ask quotes, daily share volume above 1,000 shares, daily dollar trading volume of at least $50,000, quoted spreads below 20%, and non-missing estimates of the Amihud illiquidity ratio (Amihud, 2002) and Kyle's lambda price-impact proxy (Kyle, 1985). Applying these criteria yields a final sample of 973,481 news items linked to 3,452 firms, ensuring that sentiment-based strategies are evaluated only where execution is feasible.

3 Methods

3.1 Model Families and Sentiment Extraction

The analysis compares sentiment extracted using transformer-based language models and a dictionary benchmark. Transformer models include BERT (**?**), RoBERTa (Liu et al., 2019), OPT (Zhang et al., 2022), LLaMA–3 (**?**), and FinBERT (Huang et al., 2023). BERT and RoBERTa are encoder architectures trained using masked-token prediction, whereas OPT and LLaMA–3 are decoder-only models trained using next-token prediction objectives. FinBERT adapts the BERT architecture through domain-specific pretraining on financial text. I include sentiment scores constructed from the Loughran–McDonald dictionary as a non-transformer baseline (Loughran and McDonald, 2022).

All transformer models are initialized from publicly available checkpoints released via Hugging Face and fine-tuned on Refinitiv news linked to U.S. equities. Fine-tuning adapts each model's representation layer to predict the sign of future excess returns. Feature extraction follows the probing framework of Alain and Bengio (2016), and the supervised training protocol builds on Ke et al. (2020), extended to accommodate multiple architectures and parameter scales. Model sizes are selected to remain feasible under academic compute budgets while ensuring consistent training and evaluation across architectures.

3.2 Construction of Sentiment Labels

Each news article is labeled using the associated stock's cumulative three-day excess return. The return window for an article published on day n spans $[n, n+2]$, and excess return equals the raw stock return minus the CRSP value-weighted market return. Articles receive a label of one if cumulative excess return over this window is positive and zero otherwise.

The use of a short multi-day window is consistent with event-study evidence showing that price responses to information unfold gradually due to information-processing frictions and limits to arbitrage (MacKinlay, 1997; Kothari and Warner, 2007; Mitchell and Stafford, 2000). Prior work on textual sentiment similarly documents delayed market incorporation beyond same-day returns (Tetlock, 2007b). The $[n, n+2]$ window balances timely price adjustment with robustness to noise.

Transformer models output a continuous probability $p_{i,n} \in [0,1]$ indicating the likelihood that article i on date n predicts a positive three-day excess return. These probabilities constitute the sentiment signals used in regressions and trading simulations. Dictionary-based sentiment scores are rescaled to the unit interval for comparability.

3.3 Training, Validation, and Evaluation

The filtered article sample is partitioned chronologically to prevent information leakage. Articles from January 1, 2010 to December 31, 2023 form the training set. Articles from January 1, 2024 to May 28, 2024 form the validation set. Articles from June 1, 2024 to January 30, 2026 constitute the out-of-sample test set. Split boundaries are defined so that return-label windows do not cross subsample periods.

Models are fine-tuned by minimizing cross-entropy loss between predicted probabilities and realized sentiment labels. Hyperparameters are selected based on validation-set performance. Final evaluation is conducted on the held-out test set using accuracy, precision, recall, specificity, and F1 score. These metrics characterize each model's ability to classify the sign of future excess returns using only information available at publication.

3.4 Predictive Regressions

To assess whether sentiment predicts future returns, I estimate panel regressions relating next-day stock returns to model-generated sentiment scores:

$$r_{i,n+1} = a_i + b_n + \gamma x_{i,n} + \epsilon_{i,n}, \qquad (1)$$

where $r_{i,n+1}$ denotes the return of stock i on day $n+1$. Firm fixed effects a_i absorb time-invariant heterogeneity, and date fixed effects b_n capture market-wide shocks. Standalone specifications use $x_{i,n}$ to represent the sentiment probability from a single model, allowing direct comparison across architectures. Pairwise specifications include two sentiment scores jointly to evaluate incremental explanatory content.

Standard errors are two-way clustered by firm and date to account for cross-sectional dependence and heteroscedasticity. Because all sentiment scores lie on a common probability scale, coefficient magnitudes are directly comparable across models.

3.5 Trading Framework and Execution Rules

The trading framework evaluates whether sentiment signals translate into implementable strategies under realistic market frictions. Each trading day, sentiment probabilities are merged with CRSP returns and microstructure measures, including quoted bid–ask spreads, daily trading volume, Amihud illiquidity, and standardized Kyle's lambda estimates. Liquidity screens restrict attention to stocks with reliable quotes, sufficient depth, daily share volume above 1,000 shares, daily dollar volume above $50,000, and quoted spreads below 20%.

Portfolio memberships update dynamically as new articles arrive. Stocks with at least one news item in the preceding 24 hours receive updated sentiment scores; uncovered stocks retain their previous-day signal. Each day, stocks are ranked cross-sectionally by sentiment. The highest 20% enter the long portfolio and the lowest 20% enter the short portfolio. Positions adjust only when stocks cross quintile thresholds, reducing unnecessary turnover.

All portfolios are value-weighted and self-financing. Transaction costs equal 5 basis points per trade. Participation constraints cap trade size at 10% of daily dollar volume to reflect institutional execution limits. Trades are timed according to news arrival: articles released before 6:00 a.m. generate trades at the same-day open; intraday releases trade at the close; and articles released after 4:00 p.m. trade at the next open.

This framework evaluates whether sentiment-based predictability survives realistic execution constraints and whether differences across language-model architectures translate into economically meaningful performance differentials.

3.6 Implementation Details and Reproducibility

I implement all models using the Hugging Face Transformers library with PyTorch. Each transformer is fine-tuned on the training split using cross-entropy loss with early stopping based on validation F1. I tokenize text using each model's native tokenizer and truncate inputs to the maximum supported sequence length. Evaluation protocols, return-label construction, and classification thresholds (0.50) are standardized across models.

Trading simulations reflect a backtest under explicitly stated execution assumptions. Portfolio weights update once per day based on news arriving in the preceding 24 hours, trading is restricted to the liquidity-screened universe described in Section 2, and trade sizes are capped by participation constraints tied to daily dollar volume. Transaction costs are modeled as a fixed 5 bps charge per trade. Reported results therefore measure the economic significance of model-based sentiment signals under feasible institutional execution conditions rather than frictionless theoretical returns.

4 Results

4.1 Sentiment Analysis Accuracy in U.S. Financial News

I evaluate the ability of several language models to classify the sentiment of U.S. financial news and to predict the sign of the subsequent three-day excess return. The evaluation is conducted on a held-out test set comprising 190,236 news articles that remain after applying relevance filters, novelty screens, and liquidity constraints. Each model produces a continuous probability score in the unit interval, and an article is classified as positive when the predicted probability exceeds 0.50. This probability-based framework enables a direct and architecture-agnostic comparison across models.

Table 2 reports classification accuracy, precision, recall, specificity, and F1 score for six sentiment models: BERT, RoBERTa, OPT, LLaMA–3, FinBERT, and the Loughran–McDonald dictionary. The results reveal a clear and monotonic performance ranking. LLaMA–3 achieves the strongest performance across all evaluation metrics, followed by OPT and RoBERTa. Encoder-based models such as BERT and FinBERT exhibit moderate but statistically meaningful predictive power, while the dictionary-based approach performs only slightly better than chance.

The performance ordering is stable across accuracy, recall, and F1 score, indicating that differences across models are not driven by a single metric. LLaMA–3 benefits from a larger parameter footprint and a broader pre-training corpus, which enhances its ability to generalize to het-

erogeneous financial news. OPT also performs strongly, consistent with the expressive capacity of decoder-only architectures. RoBERTa outperforms base BERT, reflecting improvements in training dynamics and contextual representation. FinBERT's domain-specific fine-tuning does not translate into superior performance, suggesting that broad contextual learning dominates narrow vocabulary specialization in this setting.

Pairwise McNemar tests on matched predictions reject equality of classification accuracy across all transformer-based models at the 1% level. These results confirm that the observed differences in predictive performance are statistically significant and not attributable to sampling variation. Overall, the evidence indicates that model scale, architecture, and training diversity play a central role in extracting predictive sentiment from financial news.

Metric	LLaMA–3	OPT	RoBERTa
Accuracy	0.782	0.763	0.748
Precision	0.766	0.751	0.739
Recall	0.801	0.776	0.762
Specificity	0.743	0.721	0.708
F1 score	0.783	0.763	0.749
Metric	BERT	FinBERT	LM Dict.
Accuracy	0.728	0.713	0.503
Precision	0.721	0.705	0.506
Recall	0.744	0.726	0.511
Specificity	0.689	0.672	0.524
F1 score	0.732	0.715	0.508

Table 2: Classification performance for sentiment models. The table reports accuracy, precision, recall, specificity, and F1 score for six models predicting the sign of the three-day excess return following each news item.

4.2 Predicting Returns with Pairwise LLM Scores

I next examine whether combining sentiment signals from multiple language models improves return predictability relative to standalone specifications. Each regression includes two sentiment scores jointly, allowing the incremental explanatory contribution of each model to be assessed while controlling for shared variation in textual information. All specifications include firm and date fixed effects, and standard errors are two-way clustered by firm and date.

Table 3 reports six representative pairwise specifications. The results reveal several clear patterns. First, combinations that include LLaMA–3 consistently produce the strongest coefficients, the highest within-R^2, and the lowest root mean squared error. When paired with either RoBERTa or OPT, the LLaMA–3 coefficient remains large and highly significant, while the accompanying model's coefficient is either attenuated or remains significant at a lower magnitude. This pattern indicates partial but incomplete overlap in the information captured by different architectures.

Second, pairings involving OPT and encoder-based models such as BERT or FinBERT also generate meaningful improvements relative to single-model regressions. These results suggest that decoder-style architectures extract sentiment dimensions that are not fully captured by masked-language models. Third, FinBERT contributes modestly in mixed specifications, but its coefficients are consistently dominated by those of OPT and LLaMA–3, reinforcing the view that broad contextual pretraining is more informative than domain-specific fine-tuning in this setting.

The dictionary-based sentiment score remains weak in all pairwise specifications and adds little incremental explanatory power once transformer-based signals are included. Improvements in R^2 and reductions in RMSE across specifications confirm that modern LLMs capture complementary aspects of sentiment that translate into stronger short-horizon return predictability when combined.

4.3 Performance of Sentiment-Based Portfolios

This subsection evaluates whether model-based sentiment delivers economically meaningful performance once mapped into implementable trading strategies. Each language model generates three value-weighted portfolios that rebalance daily: a long portfolio consisting of stocks in the top 20% of sentiment scores, a short portfolio consisting of stocks in the bottom 20%, and a zero-cost long–short portfolio that takes offsetting positions in these two groups. Value weighting aligns portfolio construction with the liquidity screens in Section 2 and reduces the influence of thinly traded small-cap names that mechanically inflate turnover.

Portfolio returns are computed net of a 5 bps one-way transaction cost per trade, intended to capture commissions and short-horizon implementation shortfall in liquid U.S. equities. Performance is measured over the June 2024–January 2026 out-

Regression	(1)	(2)	(3)	(4)	(5)	(6)
BERT score	0.11**	0.09*	0.14**			
	(2.45)	(1.98)	(2.88)			
FinBERT score	0.18***			0.21***		
	(4.12)			(4.63)		
OPT score		0.23***		0.25***		
		(5.01)		(5.38)		
RoBERTa score			0.10*		0.13**	
			(1.98)		(2.76)	
LLaMA–3 score					0.27***	0.22***
					(5.48)	(4.71)
LM dictionary score						0.07
						(1.44)
Observations	190,236	190,236	190,236	190,236	190,236	190,236
Within R^2	0.03	0.03	0.02	0.04	0.05	0.01
RMSE	4.12	4.26	4.91	3.88	3.74	9.41
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Pairwise predictive regressions including two sentiment scores jointly. The dependent variable is next-day stock return. All specifications include firm and date fixed effects. Standard errors are two-way clustered by firm and date.

of-sample window. Reported summary statistics include the Sharpe ratio, mean daily return (MDR), daily return volatility, and maximum drawdown (MDD). I use the Nasdaq Composite and the Dow Jones Industrial Average as passive buy-and-hold benchmarks over the same period.

Table 4 summarizes the sentiment-based strategy results. Portfolio performance increases monotonically with model quality. Long–short portfolios constructed from LLaMA–3, OPT, and RoBERTa sentiment scores achieve the highest Sharpe ratios and the largest average return spreads, with LLaMA–3 delivering the strongest performance. Encoder-based models such as BERT and FinBERT generate positive but materially smaller long–short returns. The dictionary-based Loughran–McDonald strategy performs weakest, exhibiting low risk-adjusted performance and substantially larger drawdowns, consistent with its limited ability to capture context-dependent tone in modern financial news.

The decomposition into long and short legs reveals an economically intuitive pattern. High-sentiment stocks earn positive subsequent returns on average, while low-sentiment stocks underperform, producing a persistent spread that is amplified for the higher-capacity models. Maximum drawdowns are notably smaller for transformer-based strategies than for the dictionary strategy, indicating greater stability in signal quality under the same execution and cost assumptions. The Dow Jones Industrial Average has a Sharpe ratio of 0.78, a mean daily return of 0.07%, daily volatility of 1.95%, and a maximum drawdown of -26.8% over the same window; the Nasdaq Composite has a Sharpe ratio of 0.88, a mean daily return of 0.09%, daily volatility of 1.98%, and a maximum drawdown of -24.2%.

4.4 Liquidity and Market Frictions

I test whether the predictive content of sentiment varies with market frictions. Liquidity shortages, wide bid–ask spreads, and price impact can slow the incorporation of news into prices and constrain investors' ability to trade on signals (Kyle, 1985; Amihud, 2002; Pastor and Stambaugh, 2003). This mechanism implies that return predictability should be stronger in securities where frictions bind more tightly.

I evaluate heterogeneity in return predictability by conditioning on market frictions. Specifically, I partition the sample using the microstructure mea-

	BERT			OPT			FinBERT		
	Long	Short	L–S	Long	Short	L–S	Long	Short	L–S
Sharpe ratio	1.30	1.12	1.95	1.55	1.33	2.45	1.18	1.05	1.75
MDR (%)	0.15	0.13	0.24	0.19	0.16	0.31	0.14	0.12	0.20
StdDev (%)	1.82	1.96	1.23	1.75	1.88	1.18	1.83	1.94	1.21
MDD (%)	-14.8	-20.4	-14.1	-12.9	-18.7	-12.5	-16.5	-22.9	-15.4
	RoBERTa			LLaMA–3			LM dictionary		
	Long	Short	L–S	Long	Short	L–S	Long	Short	L–S
Sharpe ratio	1.38	1.23	2.25	1.72	1.48	2.85	0.50	0.45	0.68
MDR (%)	0.17	0.14	0.27	0.22	0.18	0.34	0.06	0.05	0.08
StdDev (%)	1.85	1.93	1.20	1.78	1.86	1.17	2.35	2.48	1.82
MDD (%)	-16.2	-22.9	-14.7	-12.0	-17.9	-12.3	-31.2	-40.5	-34.2

Table 4: Performance statistics for value-weighted sentiment-based trading strategies. Returns incorporate a 5 bps transaction cost per trade. MDR denotes mean daily return and MDD denotes maximum drawdown. Passive benchmark statistics are reported in the text.

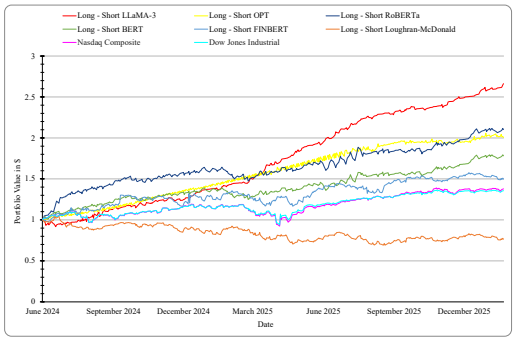

Figure 1: Cumulative returns from investing $1 in value-weighted, zero-cost long–short portfolios formed from model-specific sentiment signals. Returns are shown net of a 5 bps one-way transaction cost per trade. Nasdaq Composite and Dow Jones Industrial Average benchmarks are shown without transaction costs.

sures described in Section 2. Stocks are assigned to two groups based on a median split of the Amihud illiquidity ratio; results are qualitatively similar when using quoted bid–ask spreads. Each subsample re-estimates the baseline return-predictive regression. All specifications include firm and date fixed effects, with standard errors clustered two-way by firm and date. The liquidity-split regressions use all observations with non-missing return, sentiment, and microstructure variables, whereas the classification results in Section 4.1 are reported on the held-out test period only.

Table 5 reports the results. Predictive coefficients increase monotonically from the high-liquidity bucket to the low-liquidity bucket for all transformer-based models, with the strongest liquidity gradient for decoder models (LLaMA–3 and OPT). This pattern is consistent with richer sentiment representations taking longer to be incorporated into prices when trading frictions limit immediate arbitrage. Encoder models (RoBERTa, BERT, and FinBERT) also exhibit a positive liquidity gradient, but with smaller magnitudes. The dictionary-based signal remains weak in both buckets, indicating that lexicon scores do not reliably capture the components of news that diffuse slowly under market frictions.

5 Conclusion

This paper evaluates whether large language models extract return-relevant information from firm-specific news that can be converted into implementable portfolio signals under realistic market frictions. The empirical design combines three complementary layers of evidence: sentiment classification for three-day excess returns, predictive regressions for next-day returns with firm and date fixed effects, and value-weighted trading strategies that incorporate transaction costs, liquidity constraints, and execution timing tied to news arrival.

Modern transformer-based language models consistently outperform traditional lexicon-based sentiment. Classification tests show that LLaMA–3 achieves the strongest accuracy, precision, recall, and F1 scores, followed by OPT and RoBERTa, while BERT and FinBERT deliver moderate predictive performance. The Loughran–McDonald dictionary performs close to chance. This ordering persists in predictive regressions. Decoder-style architectures such as LLaMA–3 and OPT gener-

	High liquidity (low frictions)			Low liquidity (high frictions)		
Model	Coef.	t-stat	N	Coef.	t-stat	N
LLaMA–3	0.24***	(5.10)	406,271	0.38***	(7.46)	406,272
OPT	0.20***	(4.72)	406,271	0.36***	(7.08)	406,272
RoBERTa	0.16***	(3.58)	406,271	0.26***	(5.62)	406,272
BERT	0.10**	(2.64)	406,271	0.18***	(4.55)	406,272
FinBERT	0.13***	(3.41)	406,271	0.23***	(5.21)	406,272
LM Dict.	0.05	(1.33)	406,271	0.09*	(1.98)	406,272
Firm FE	Yes			Yes		
Date FE	Yes			Yes		

Table 5: Predictive regression estimates by liquidity bucket. The dependent variable is next-day stock return. Liquidity buckets are formed using a median split of the Amihud illiquidity ratio computed in Section 2. All specifications include firm and date fixed effects with two-way clustered standard errors.

ate substantially larger predictive coefficients than encoder baselines, whereas the dictionary-based signal exhibits weak and economically small explanatory power. Pairwise specifications further show that high-capacity models retain incremental predictive content when combined with other transformers, indicating that they capture sentiment dimensions not absorbed by simpler representations.

Trading simulations show that these informational differences translate into economically meaningful performance once realistic frictions are imposed. Value-weighted, zero-cost long–short portfolios formed from LLaMA–3, OPT, and RoBERTa sentiment scores outperform both market benchmarks and the dictionary strategy during the June 2024–January 2026 out-of-sample period after accounting for transaction costs and participation constraints. The LLaMA–3 long–short portfolio attains a cumulative return of approximately 180% with a Sharpe ratio of 2.85, followed by OPT at roughly 155% (Sharpe 2.45) and RoBERTa at roughly 120% (Sharpe 2.25). BERT and FinBERT also generate positive long–short spreads, while the dictionary strategy ends the period below the initial investment value. These results indicate that LLM-derived sentiment contains richer and more actionable information than lexicon methods and that the resulting signals remain economically meaningful under feasible execution assumptions.

The paper's primary contribution is to position LLM sentiment as a *portfolio input* rather than a descriptive text measure. Prior research has established that text can forecast returns, but evidence is thinner on whether LLM-based signals remain economically valuable once mapped into explicit trading rules that respect liquidity, trading capacity, and timing constraints. The results show that probability outputs from modern language models can be operationalized for portfolio construction and evaluated in a forward-looking setting where implementation feasibility matters.

The findings have implications for practitioners and regulators. Asset managers should view LLM-based news sentiment as a complement to existing signal libraries when portfolio construction explicitly accounts for liquidity and execution frictions. Regulators and market designers should recognize that LLM adoption in trading and risk analytics may affect how quickly public information is processed and incorporated into prices, raising questions about market stability, liquidity provision, and the distribution of informational advantages.

Several avenues for future research remain. Extending the framework to intraday horizons with richer execution models would clarify how signal value depends on latency and order-book conditions. Combining news sentiment with other unstructured sources such as earnings calls, regulatory filings, or social media may improve robustness and reduce dependence on any single channel. Developing interpretable LLM pipelines that identify which textual elements drive trading decisions would further narrow the gap between predictive performance and economically grounded explanations.

6 Limitations

Several limitations remain.

First, the analysis focuses exclusively on U.S. equities and English-language financial news. While

this choice ensures data quality and execution feasibility, it limits the generalizability of the findings to other markets, asset classes, and languages. Sentiment dynamics, liquidity conditions, and information diffusion may differ substantially in emerging markets or non-English news environments.

Second, although the trading simulations incorporate realistic transaction costs, participation limits, and liquidity screens, they remain an abstraction of real-world execution. The analysis does not model intraday order-book dynamics, strategic interaction among traders, or endogenous market impact arising from widespread adoption of similar signals. Realized performance may therefore differ if such strategies are deployed at scale.

Third, the use of large language models introduces opacity and interpretability challenges. While the models demonstrate strong predictive performance, the specific linguistic features driving sentiment signals are not fully transparent. This limits economic interpretability and complicates attribution of predictions to particular textual mechanisms.

Finally, the findings raise potential risks related to unequal access to advanced language models and computational resources. Widespread use of high-capacity LLMs in trading could advantage well-resourced institutions and may affect market efficiency, liquidity provision, and price discovery. These broader systemic implications are beyond the scope of this paper but warrant careful consideration in future research.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. ArXiv:1610.01644.

Yakov Amihud. 2002. Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56.

Hailiang Chen, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.

Zhi Da, Joseph Engelberg, and Pengjie Gao. 2011. In search of attention. *Journal of Finance*, 66(5):1461–1499.

Allen H. Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Andrew H Huang, Siew Hong Teoh, and Y Zhang. 2014. Tone management. *Review of Financial Studies*, 27(3):1043–1083.

Narasimhan Jegadeesh and Di Wu. 2013. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Zheng Ke, Bryan T. Kelly, and Dacheng Xiu. 2020. Predicting returns with text data. SSRN 3389884.

S. P. Kothari and Jerold B. Warner. 2007. Econometrics of event studies. In B. Espen Eckbo, editor, *Handbook of Corporate Finance: Empirical Corporate Finance*, pages 3–36. Elsevier.

Albert S. Kyle. 1985. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Feng Li. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3):221–247.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

TIM Loughran and BILL Mcdonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65.

Tim Loughran and Bill McDonald. 2022. Master Loughran-MacDonald Word Dictionary.

A. C. MacKinlay. 1997. Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Asaf Manela and Alan Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Mark L. Mitchell and Erik Stafford. 2000. Managerial decisions and long-term stock price performance. *Journal of Business*, 73(3):287–329.

Lüboš Pastor and Robert F. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685.

S. Michael Price, James S. Doran, David R. Peterson, and Brian A. Bliss. 2015. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Financial Economics*, 115(3):415–430.

Paul C Tetlock. 2007a. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Paul C. Tetlock. 2007b. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3):1437–1467.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.