

News Text Embeddings and Clean Energy ETF Return Direction Prediction: An Empirical Study Based on BERT-PCA and LightGBM

Weiliang Liu

Universiti Putra Malaysia, Selangor, 47300, Malaysia

ABSTRACT

This study investigates whether deep semantic information from financial news headlines provides incremental predictive value for the weekly price direction of the iShares Global Clean Energy ETF (ICLN). Using 189 weeks of data from June 2020 to March 2024, we construct a predictive framework combining BERT embeddings, PCA dimensionality reduction, and a LightGBM classifier, evaluated under a rigorous rolling window validation protocol. Empirical results reveal that although textual factors dominate feature importance—contributing 92.2% of total gain—the full model achieves an out-of-sample AUC of only 0.5359 and a recall of just 11% for upward weeks, failing to generate effective trading signals. These findings indicate that, under the current sample and model specifications, news headline semantics do not provide stable predictive power for ICLN's weekly direction. Methodologically, this paper contributes a transparent pipeline for transforming unstructured news into structured predictors and establishes a rigorous evaluation framework suitable for small-sample financial time series, offering a valuable reference for empirical research in financial natural language processing.

KEYWORDS

Text Embeddings; BERT; Clean energy ETF; Return prediction; LightGBM

1 Introduction

The rapid proliferation of thematic exchange-traded funds (ETFs) has heightened investor demand for timely signals capable of predicting short-term price movements. Among these products, the iShares Global Clean Energy ETF (ICLN) has consistently attracted attention, as its constituent stocks are influenced not only by firm-level fundamentals but also by macroeconomic conditions, commodity prices, policy announcements, and shifts in investor sentiment regarding the energy transition. Consequently, ICLN's return dynamics are particularly sensitive to information shocks that first manifest through news flow, making it an ideal setting for evaluating whether modern natural language processing (NLP) text representations—beyond traditional sentiment scores—contain incremental predictive information for near-term returns.

Predicting asset returns, especially at a weekly frequency, remains challenging. In efficient markets, publicly available information is rapidly incorporated into prices, leaving limited exploitable predictability after accounting for risk factors and transaction costs. Nevertheless, a growing body of literature suggests that text-based information can predict market activity and returns in specific contexts, including earnings announcements, macroeconomic releases, and sector- or theme-specific narratives. Early approaches typically relied on dictionary-based sentiment measures or manually engineered textual features, which can be fragile as they often overlook context, sarcasm, negation, and domain-specific meanings prevalent in financial language. Jiang et al. (2021), in an empirical study on Chinese financial markets, demonstrate that media textual sentiment constructed using a finance-specific sentiment dictionary significantly predicts stock returns with robust out-of-sample performance, illustrating the generalizability of textual predictive power across different market environments.

Transformer-based language models address these limitations by learning contextual embeddings that map text into dense vectors, providing richer semantic representations and enabling downstream predictive models to exploit nonlinear interactions between news and market variables. Jiang, Liu, and Meng (2024) further show that sentiment measures derived from large language models outperform traditional dictionary methods in predicting market returns, particularly during economic downturns and extreme risk events. Their findings provide strong empirical support for adopting BERT embeddings in this study.

This research investigates whether weekly aggregated BERT embeddings extracted from financial news headlines can improve predictions of ICLN's price direction for the following week. The empirical design emphasizes practicality: the available sample is relatively small (approximately three years of weekly observations), which is typical for thematic ETFs and imposes constraints on high-capacity deep learning models. To preserve the advantages of contextual text representations while mitigating overfitting, we combine (i) Transformer embeddings as high-dimensional news features, (ii) Principal Component Analysis (PCA) for dimensionality reduction, and (iii) a gradient boosting decision tree classifier (LightGBM) trained on lagged weekly predictors. The model design strictly follows information timing: predictors from week t (including news aggregated within week t , market control variables, and lagged variables) are used to forecast the sign of the log return in week $t + 1$.

This paper makes three contributions. First, it constructs a weekly “news factor” for a clean energy ETF by averaging headline-level BERT embeddings per week, offering a transparent and replicable mechanism for mapping unstructured news into structured predictors. Second, it proposes a parsimonious modeling framework suitable for small-sample financial time series: PCA compresses the high-dimensional embedding space into a limited set of latent components, while LightGBM flexibly captures nonlinearities and interactions between news-derived signals and standard market covariates (e.g., ICLN volume, S&P 500 returns, and crude oil price changes). This approach aligns with the methodology systematically outlined in Jiang, Tang, and Ma (2024), which provides the theoretical foundation for combining dimensionality reduction with tree-based models. Third, it evaluates predictive performance using temporally appropriate validation (rolling window evaluation) and reports discrimination metrics such as the area under the ROC curve (AUC), which is robust to class imbalance and well-suited for directional prediction tasks.

Overall, the goal is not to assert the existence of persistent arbitrage opportunities, but to quantify whether contextual news information provides measurable incremental predictability for ICLN direction after imposing standard predictors and strict timing rules.

2 Related Literature

2.1 Textual Information and Asset Price Dynamics

A longstanding strand of financial economics examines how public information is incorporated into prices and trading activity. Early evidence indicates that news content is not merely a contemporaneous reflection of market conditions but may carry incremental predictive information. Tetlock (2007) demonstrates that the tone of news content correlates with subsequent market outcomes and trading activity, supporting the premise that textual signals can be transformed into measurable predictors. Engelberg and Parsons (2011) provide evidence that media-driven information diffusion affects investor attention and trading behavior, highlighting mechanisms through which news can drive prices rather than merely report them. Recent developments in Chinese financial markets further confirm these patterns: Jiang et al. (2021) find that media textual sentiment constructed using a finance-specific Chinese sentiment dictionary significantly predicts stock returns both in-sample and out-of-sample, demonstrating that textual predictive ability extends beyond Western markets. Jiang, Liu, and Meng (2024) extend this line of inquiry by showing that large language model-based sentiment measures outperform traditional dictionary methods in predicting market returns, particularly during economic downturns and extreme risk events. Collectively, these studies provide a solid empirical foundation for adopting BERT embeddings in this research.

2.2 Finance-Specific Language and Measurement Issues

A key methodological challenge is that financial language differs substantially from general-purpose sentiment dictionaries, creating risks of misclassification when simple sentiment tools are applied. Loughran and McDonald (2011) formalize this issue by developing a finance-oriented word list and documenting the systematic errors that general-purpose dictionaries produce in financial contexts. Loughran and McDonald (2020) synthesize best practices and pitfalls in financial text analysis—particularly concerning construct validity, look-ahead bias, and evaluation design. Parallel developments in the Chinese context address similar issues. Jiang et al. (2021) develop a comprehensive Chinese financial sentiment dictionary, demonstrating that general-purpose sentiment dictionaries generate substantial measurement errors in financial contexts. This domain-specific dictionary construction approach motivates the use of contextual representations that can better capture financial semantics.

2.3 Contextual Embeddings and Financial Domain Representations

Recent advances in NLP have provided practical means to represent text with richer semantics. Devlin et al. (2019) introduce BERT, demonstrating that deep contextual representations learned via the Transformer architecture significantly improve downstream tasks relative to static embeddings. For financial applications, domain adaptation is particularly important: Araci (2019) proposes FinBERT, showing that pre-training on financial corpora helps align embeddings with domain-specific usages. Recent work by Jiang, Liu, and Meng (2024) trains a Chinese financial large language model that integrates structured market data with unstructured text, achieving performance superior to traditional methods in sentiment measurement and risk prediction. Together, these studies support the use of embedding-based features to characterize weekly news flow, capturing latent semantic signals beyond coarse sentiment.

2.4 Machine Learning and Return Prediction

Machine learning has become central to modeling nonlinear relationships between predictors and asset returns. Gu, Kelly, and Xiu (2020) provide a comprehensive empirical asset pricing framework, emphasizing that flexible ML models can extract predictive structure from large sets of covariates while highlighting the necessity of rigorous out-of-sample testing. In tabular prediction settings, gradient boosting decision tree (GBDT) methods are widely adopted due to their strong performance, robustness to feature scaling, and ability to model nonlinear interactions. Ke et al. (2017) propose

LightGBM, an efficient GBDT implementation that scales well and incorporates regularization and sampling strategies. Jiang, Tang, and Ma (2024) systematically cover dimensionality reduction, tree models, and text analysis techniques, providing direct reference for the methodological choices in this paper. Their recent research on multi-feature machine learning (Jiang et al., 2025) further demonstrates how ML frameworks can uncover nonlinear pricing effects in asset markets, offering theoretical support for introducing nonlinear interactions in LightGBM.

2.5 News Uncertainty and Macro-Financial Variables

Beyond firm- or sector-specific headlines, macroeconomic uncertainty and commodity/market benchmarks can affect clean energy assets through discount rate, growth, and risk premium channels. Baker, Bloom, and Davis (2016) develop a news-based measure of economic policy uncertainty, providing conceptual justification for including broader market and commodity covariates alongside textual features. For a clean energy ETF like ICLN, such control variables are particularly reasonable given its sensitivity to energy prices, growth expectations, and risk sentiment.

2.6 Implications for This Study

Taken together, these research streams suggest a coherent modeling strategy: construct domain-relevant textual predictors using contextual embeddings; combine text-derived signals with market and macroeconomic covariates; and estimate the predictive mapping using robust tabular ML baselines such as LightGBM with strict time-series validation. Following these principles, this study uses lagged weekly features extracted from BERT embeddings and financial market variables to predict the following week's price direction of ICLN.

3 Data

3.1 Sample Construction and Frequency

The empirical analysis is conducted at a weekly frequency to align market data with the flow of information in news text and to mitigate microstructure noise present in high-frequency returns. The sample covers approximately three years of weekly observations (June 2020 to March 2024), totaling 189 weeks, resulting in a relatively small panel suitable for parsimonious, regularized modeling choices. Weeks are defined using a consistent weekend convention (Friday closing prices). Market variables are aligned to the same weekly date index before feature engineering and modeling.

3.2 Market Variables

The prediction target is constructed from weekly returns of ICLN. Specifically, weekly log returns are calculated from adjusted prices, and the dependent variable is a binary direction indicator:

$$y_t = \mathbb{1}(r_{t+1} > 0)$$

where P_t denotes the adjusted price of ICLN at week t . In addition to ICLN's own lagged return dynamics, the model includes lagged weekly returns of a broad equity benchmark (S&P 500) and an energy price proxy (WTI crude oil). Weekly trading activity in ICLN is captured by weekly volume, defined as the sum of daily volumes within each week. All market covariates are used strictly in lagged form to avoid look-ahead bias.

3.3 Textual Data and BERT Embeddings

The textual information comes from a news headline dataset. Headlines are embedded using a pre-trained BERT model to obtain dense semantic representations. Embeddings are aggregated in two stages: (1) daily aggregation—embeddings for the same calendar date are averaged element-wise; (2) weekly aggregation—daily embeddings within the same week are averaged element-wise to form a single weekly embedding vector. Weeks without valid embeddings are excluded.

3.4 Dimensionality Reduction via PCA

Raw BERT embeddings are high-dimensional (768 dimensions) and may be noisy in small samples. To improve statistical efficiency and reduce overfitting risk, the weekly embedding vectors are transformed using PCA. PCA is fitted only on the training sample of each modeling split to prevent information leakage. The number of retained components is treated as a tunable hyperparameter.

3.5 Feature Engineering and Final Modeling Table

The final modeling dataset is constructed using two lags of: (i) PCA-compressed textual features, (ii) ICLN returns and volume, and (iii) S&P 500 and WTI returns. By construction, all predictors at time t are functions of information available no later than week t , ensuring that the classification task can be implemented in real time without look-ahead leakage.

4 Methodology

4.1 Problem Setup and Notation

Let t index weeks ending on Friday. Let P_t denote the adjusted closing price of ICLN in week t , and define the weekly

log return as $r_t = \log(P_t) - \log(P_{t-1})$. The prediction target is $y_t = \mathbb{1}\{r_{t+1} > 0\}$, meaning the model uses information available up to week t to predict whether the next week's return will be positive. The feature vector x_t consists of two weekly lags of ICLN returns, ICLN volume, S&P 500 returns, WTI returns, and BERT embeddings.

4.2 Data Alignment, Aggregation, and Lag Construction

All series are aligned to a common weekly calendar. Market series are constructed from daily data using the last available adjusted closing price within each week for prices and the sum of daily volumes for volume. For textual data, embeddings within each week are averaged to obtain e_t . Weeks without headlines are dropped. Lag terms are generated deterministically using time shifts to eliminate look-ahead bias.

4.3 Dimensionality Reduction of BERT Embeddings via PCA

Raw BERT embeddings are compressed using PCA after standardization. Leakage prevention is critical: standardization parameters and PCA loadings are estimated using only the training sample. For each split, the transformed embeddings are computed as $u_t = \Pi_K(S(e_t))$, and the same fitted transformation is applied to validation data.

4.4 Prediction Model: LightGBM

The core prediction model is a LightGBM classifier. LightGBM is well-suited for this application because it performs strongly on small- to medium-sized tabular datasets, incorporates explicit regularization controls, and supports early stopping.

4.5 Time-Series Validation and Early Stopping

The study employs rolling window validation: a fixed training window of 130 weeks, yielding 59 out-of-sample predictions. At each rolling step, PCA and the standardizer are fitted solely on training data, and LightGBM is trained with early stopping using a validation set (20% of training data).

4.6 Hyperparameter Optimization with Optuna

Hyperparameter selection is performed using Bayesian optimization via Optuna, maximizing validation AUC. The search space covers PCA components, learning rate, tree structure parameters, subsampling controls, and regularization strengths. The optimal parameters are selected in the first training window and fixed thereafter.

4.7 Performance Metrics and Reporting

The primary metric is AUC. For interpretability, threshold-based metrics (accuracy, precision, recall, F1) are reported at a threshold of 0.5.

5 Empirical Results

5.1 Evaluation Design and Information Set Alignment

Rolling window validation: total 189 weeks, training window 130 weeks, 59 out-of-sample predictions. The proportion of positive weeks is 45.8%, with no significant class imbalance.

5.2 Textual Feature Dimensionality Reduction (PCA)

With $K = 27$, the average cumulative explained variance across rolling windows is 73.9%.

5.3 Hyperparameter Optimization and Selected Model

Optuna selects: $K = 27$, learning_rate ≈ 0.06 , max_depth = 2, and other parameters as detailed in the full paper.

5.4 Predictive Performance and Benchmark Comparison

The majority-class benchmark yields an AUC of 0.5; the LightGBM model using only financial features achieves an AUC of 0.4491; the complete model incorporating BERT-PCA features achieves an AUC of 0.5359. At a threshold of 0.5, accuracy is 49.15%, with recall of only 11% for upward weeks.

Table 1 Rolling Window Validation AUC Comparison

Model	AUC
Majority Class Benchmark	0.5000
LightGBM (Financial Features Only)	0.4491
LightGBM (Financial + BERT-PCA)	0.5359

Table 2 Complete Model Confusion Matrix (Threshold 0.5)

Actual \ Predicted	Predicted 0	Predicted 1
Actual 0 (Down)	26	6
Actual 1 (Up)	24	3

5.5 Economic Value Assessment

Given the low AUC, the model does not provide practical trading signals; economic value assessment is omitted.

5.6 Model Interpretability and Textual Information Contribution

Textual factors contribute 92.2% of total gain importance. Permutation importance shows that shuffling text factors reduces AUC to approximately 0.45, indicating that textual information is the primary source of predictions but its direction is weakly associated with true labels.

5.7 Robustness and Sensitivity Analysis

Varying PCA dimensions ($K = 10\text{--}40$) yields AUCs between 0.512 and 0.540. Changing lag orders (1 or 3) yields AUCs of 0.503 and 0.518. Adjusting training window length (104–156 weeks) yields AUCs of 0.521–0.543. Excluding the 2022 energy crisis period yields an AUC of 0.542. Overall, predictive ability remains weak.

6 Discussion and Limitations

The BERT-PCA + LightGBM framework achieves an AUC of only 0.5359, significantly lower than expected, indicating that deep semantic information from news headlines does not provide stable predictive power for ICLN's weekly direction. The model's extremely low recall for upward weeks (11%) further confirms its limited practical value.

Several factors may explain this outcome: limited sample size (189 weeks); fixed hyperparameters across rolling windows; noisy headlines with weak associations to ETF price movements; weak predictive power of market features; and potential information loss from PCA compression.

Methodologically, this study contributes a replicable rolling validation framework and highlights the challenges of model evaluation with small samples in financial NLP research. Future work could explore expanding the sample, using domain-adapted models such as FinBERT, implementing dynamic hyperparameter tuning, or integrating additional data sources.

7 Conclusion

This study investigates whether deep semantic information from news headlines can provide incremental predictive value for the weekly price direction of ICLN. Using a BERT-PCA + LightGBM framework with rolling window validation on 189 weeks of data, we find that although textual factors dominate feature importance (92.2% gain), the full model achieves an out-of-sample AUC of only 0.5359 and a recall of 11% for upward weeks. Compared to the financial-only benchmark (AUC 0.4491), incorporating BERT embeddings yields a relative improvement of approximately 19%, but absolute predictive power remains weak and fails to deliver effective trading signals.

Despite limited predictive performance, this study makes methodological contributions. It provides a transparent pipeline for converting unstructured news into structured predictors and establishes a rigorous evaluation framework for small-sample financial time series, emphasizing rolling window validation, prevention of look-ahead bias, and comprehensive use of multiple evaluation metrics.

Limitations point to future directions: expanding sample size, using finance-specific models like FinBERT, implementing dynamic hyperparameter tuning, and integrating richer textual sources. Although this study does not confirm stable predictive power from news embeddings, its methodological framework offers a useful reference for empirical research in financial NLP.

References

- [1] Jiang F., Meng L., & Tang G. (2021). Media Textual Sentiment and Stock Return Prediction. *Economics (Quarterly)*, (04), 1323-1344.
- [2] Jiang F., Liu Y., & Meng L. (2024). Large Language Models, Textual Sentiment, and Financial Markets. *Management World*, (8).
- [3] Jiang F., Tang G., & Ma T. (2024). *Machine Learning for Finance and Financial Management*. Beijing: China Machine Press.
- [4] Jiang F., Tang G., Chen H., & Zhu L. (2025). Asset Pricing with Investor Expectation Coordination. *Management World*, (12).
- [5] Araci D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
- [6] Baker S.R., Bloom N., & Davis S.J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636.
- [7] Devlin J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT 2019**.
- [8] Engelberg J.E., & Parsons C.A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67-97.
- [9] Gu S., Kelly B., & Xiu D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.
- [10] Ke G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS 2017*.
- [11] Loughran T., & McDonald B. (2011). When is a liability not a liability? *The Journal of Finance*, 66(1), 35-65.
- [12] Loughran T., & McDonald B. (2020). Textual analysis in accounting and finance. *The Journal of Finance*, 75(3), 1187-1230.
- [13] Tetlock P.C. (2007). Giving content to investor sentiment. *The Journal of Finance*, 62(3), 1139-1168.