Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

Analyzing the Impact of Social Media Sentiment on Stock Investment Decisions Using Machine Learning Techniques

Dr. Muskan[1], Dr. Pankaj Kumar[2]

[1]Assistant Professor, Faculty of Management, SRM IST, Ghaziabad, Uttar Pradesh

[2]Assistant Professor, Faculty of Management, SRM IST, Ghaziabad, Uttar Pradesh

Email: mrani8002@gmail.com[1], pankaj.manpur1992@gmail.com[2]

Abstract

The study of social media analytics and financial market analysis has become one of the most important areas of current computational finance research. In this paper, we perform an extensive empirical analysis on the potential for machine learning (ML) approaches to capture and leverage sentiment information from social media to improve the process of investing in stocks. By using data over the course of several years, between 2021 and 2024, collected from more than 48 million social media posts, news articles, and financial analyst reports from social media platforms such as X (formerly Twitter), Reddit (r/WallStreetBets and r/investing), StockTwits, and news feeds, we construct and validate an ensemble method based on transformer architectures for NLP tasks along with LSTM and gradient-boosted ML algorithms. In our approach, we consider sentiment polarity scores, topic modeling, named entity recognition, as well as time-based aggregation to create sentiment feature vectors combined with conventional financial data, such as price/volume metrics, measures of volatility, macroeconomic signals, and momentum features at a sector level. Empirical findings from applying ML models to portfolios of 200 S&P 500 stocks show that our sentiment-based approach is capable of delivering statistical significance in generating alpha of 8.3% annually, beating benchmarks by Sharpe ratio by 0.41 points relative to purely technical approaches. We explore underlying mechanisms, consider the differential informativeness of retail vs. institutional investors' discussion on social media, as well as study dynamics of cascades as a tool for intensifying impact of sentiment shocks. Our results have important practical applications in designing trading algorithms, understanding retail behavior, and regulating social media market manipulations. We discuss the limitations of our study carefully.

Keywords: Sentiment analysis of social media posts, stock market forecasting, machine learning, natural language processing, FinBERT model, LSTM neural network, transformer architecture, automated stock trading, retail investors' behavior, emotion-driven investments, financial natural language processing

1. Introduction

Financial markets are fundamentally informational systems, and the emergence of social media technology has heralded an information processing revolution in terms of volume, frequency, and diversity of information streams. Currently, Twitter/X, Reddit, and StockTwits collectively provide billions of data points pertaining to the financial markets each day, creating an informational system that traditional quantitative models were simply not constructed to handle. Indeed, the events of 2021 surrounding the emergence of meme stocks, most notable amongst which was the GameStop (GME)/AMC situation, provided clear evidence that the information processing system created through social media could create systemic market dislocations, and thus regulators, institutions, and researchers started taking an intense interest in quantitative modeling of investor sentiment online (Anand & Pathak, 2022; Barberis et al., 2023).

While literature regarding the predictive power of news sentiment in terms of returns on equity dates back to the pioneering research into media and financial markets in the early 2000s, the current age of social media calls for new methods of analysis. Social media, contrary to financial news, features extreme linguistic variance, including abbreviations, use of slang, and sarcastic or ironic comments, among other peculiarities. Indeed, the development of massive pre-trained language models, followed by their fine-tuning on financial text streams (Shah et al., 2022), opens up novel avenues for accurate identification of the sentiment signal latent in the noisy stream. At the same time, progress in deep learning architectures – particularly in attention-based methods and transformer models – makes possible the identification of contextual, relational, and temporal features in the stream of sequential financial text that could not be captured before (Li et al., 2024).

Despite rapid growth in the literature in this area, certain limitations persist. Specifically, most existing studies concentrate exclusively on a small set of platforms (most prominently Twitter) and a select subset of stocks, thus

limiting their generalizability. Additionally, a lack of separation between contributions of sentiment, macroeconomics, and market microstructure indicators renders the isolation of the incremental role of social media data problematic. Furthermore, there is a gap in translating scientific insights into trading strategies that continue to be profitable after transaction costs and latency effects are taken into account. Lastly, a lack of consideration of the ethical implications of using sentiment for trading purposes persists, even though these can contribute to increased volatility and manipulation, among other issues.

These deficiencies have been addressed in this paper through an extensive, multi-channel empirical analysis conducted over 2021-2024. Our contributions include the following: (1) Development of a multi-stage machine learning approach that combines NLP using transformers, time-series aggregation, and multi-factor fusion for stock price prediction based on sentiments; (2) Comprehensive out-of-sample testing of our models in varied market environments, such as those of volatility witnessed in 2022 and AI-generated bull markets of 2023-2024; (3) A breakdown of sources of sentiment signals on various platforms, across different classes of assets and types of investors; and (4) An exploration of ethical and regulatory aspects of sentiment-based trading.

1.1 Research Objectives

The research aims to address the following five core research questions: (i) to explore the added value of social media sentiment variables as predictors of next-day and next-week stock returns in comparison with conventional quantitative variables; (ii) to test different ML architectures, such as conventional machine learning, RNNs, and transformers, for the extraction of financial sentiment and return predictions; (iii) to identify platform-specific differences with respect to the usefulness of sentiment information and timing properties; (iv) to backtest various investment strategies that incorporate financial sentiment signals; and (v) to investigate the behavioral and structural mechanisms behind sentiment-driven price formation.

1.2 Scope and Significance

Significance of Our Work

This research is significant from various perspectives. The insight that our paper offers can be useful for asset managers who use alternative data in their investment strategies. For individual investors, we identify factors through which their social media activity impacts their decisions. For regulatory bodies such as SEC and ESMA, our paper highlights the dynamics of sentiment manipulation and asymmetry of information that can aid in regulatory policy formation. For scholars who conduct research on computational finance and financial NLP, our methodology provides a strong basis for benchmarking and identification going forward.

2. Literature Review

2.1 Evolution of Sentiment Analysis in Finance

Textual sentiment analysis for the finance field has an extensive and ever-growing academic literature base. Initial research employed dictionary methods like the Harvard General Inquirer and Loughran-McDonald Financial Sentiment Word List, to measure sentiment expressed by annual reports, earnings calls, and financial news stories (Loughran & McDonald, 2011). Though these lexicon-based approaches displayed significant predictive accuracy for returns and volatility in the short term, the fact that they did not accommodate for context, negation, and specialized vocabulary set up evident restrictions, especially concerning social media content (Kumar & Ravi, 2022).

The advent of supervised machine learning techniques such as SVM, Naive Bayes classifier, and random forests for financial sentiment analysis provided a major breakthrough in terms of modeling capability, as financial sentiment models could learn domain-specific sentiment through labeled datasets. Ding et al. (2021) showed that event-based deep learning models using financial news could forecast stock price movements with over 60% accuracy, far better than lexicon baseline models. Pre-training transformers using BERT architecture (Devlin et al., 2019) and its domain-specific variation, FinBERT (Araci, 2019), marked a new era in financial NLP, outperforming earlier methods in sentiment analysis, named entity recognition, and relation extraction tasks.

The recent literature has shifted its focus towards exploiting the unique characteristics of data from social media for financial forecasting purposes. The study carried out by Chen et al. (2022) showed that there was predictive value for short options market activity in Reddit post activity from the forum r/WallStreetBets in the context of the meme stocks event in 2021. Previous pioneering research on predicting the behavior of the Dow Jones Index based on tweets using mood analysis conducted by Bollen et al. indicated that the state of collective emotions especially when they are calm and happy could predict future market movements with 87.6% precision. Recent studies have considered multi-modal signals such as network structure and trading volume to name a few.

2.2 Machine Learning Methods for Market Prediction

The use of machine learning in the context of stock market predictions covers a vast number of methods and techniques. Classical supervised learning approaches including GBM, random forest algorithms, and regularized linear models remain solid benchmarks for financial prediction in tabular data because of their high resistance to overfitting (Gu et al., 2021). On the other hand, deep neural networks such as LSTM and GRU have proven to be especially effective in extracting temporal information from financial time series, owing to their inherent ability to capture non-linear lags without manual feature extraction (Fischer & Krauss, 2022).

The introduction of transformers in natural language processing for financial applications has brought a revolutionary change into the industry through introducing contextual embeddings in financial text that capture the semantic relationship between sentences in large-scale documents (Wu et al., 2023; Zhang et al., 2024). Transformer-based models including FinBERT, FinGPT, BloombergGPT, and even domain-specific fine-tuning models outperform other approaches on financial sentiment prediction benchmarks. Reinforcement learning approaches have been used in portfolio optimization settings, with policy gradient models proving convergence to the optimal solution under different levels of sentiment uncertainty (Xiong et al., 2023). The concept of graph neural networks is gaining prominence as an effective approach to represent inter-firm information spillovers within social networks and modeling the network-wide transmission of sentiment shocks between related firms (Feng et al., 2022).

There is empirical evidence that ensemble approaches comprising heterogeneous machine learning models significantly outperform any single architectural solution when applied to financial forecasting problems. Hybrid architectural solutions that incorporate sentiment embeddings based on transformers into technical analysis via gradient boosting modules have emerged as the de facto benchmark for sentiment-driven quantitative trading systems (Huang et al., 2024). The problem of eliminating lookahead bias and information leakage within financial ML workflows has gained increasing attention among professionals and led to the development of robust walk-forward validation techniques tailored for financial time series (Bailey et al., 2022).

2.3 Behavioral Finance and Social Influence

Underlying social media's effects on the stock market via sentiment analysis is the framework provided by behavioral finance theory. As Shiller (2019) puts forward in his theory of narrative economics, public narratives circulate among social networks and shape economic behavior, leading to observable economic outcomes in the market. Investor attention theory postulated by Barber and Odean (2008) and its adaptation to the social media environment through Seasholes and Wu (2021) indicates that investors tend to follow attention-generating securities – those discussed in social media – thereby causing herding, momentum, and ultimately mean-reversion behavior.

Financial contagion theory as a phenomenon where sentiment states propagate through investor networks and coordinate behavior has been empirically shown in cases of Reddit, Twitter, and financial blogs (Ramos et al., 2023; Chen et al., 2022). In line with information cascade models from theoretical sociology, social media's impact on stock prices has been studied in an attempt to describe when and why herding amplifies information beyond fundamentals in its price impact (Guo et al., 2021). Finally, survey data and laboratory experiments indicate a strong relationship between investor social media usage and decisions about investments (Banerjee et al., 2023).

2.4 Alternative Data in Systematic Investing

A systematic integration of alternative data, or any data sources apart from financial statements and market data, is a characteristic feature of current-day quantitative investing. The types of alternative data being used most actively by hedge funds and systematic investors include satellite imagery, credit card transactions data, web analytics, and social media sentiment data (Kumar & Ravi, 2022). Industry surveys show that as of 2023, over 70% of large quantitative funds make use of at least one type of alternative data sources, while social media sentiment is amongst the top three (Li et al., 2024).

A significant expansion in academic research related to alternative data sources in finance occurred after 2021 as evidenced by a series of studies analyzing incremental alphas obtained through web scraping of job openings, patents, ESG social media signals, and congressional disclosures (Yadav et al., 2023; Fischer & Krauss, 2022). The problem of excess returns achieved through access to data provided by the vendor followed by the dilution effect due to a crowded signal is particularly relevant for social media sentiment, gaining increasing attention since 2021.

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

3. Data and Preprocessing

3.1 Social Media Data Collection

The main methodology employed for gathering data relied on a variety of means including APIs provided by organizations themselves, web scraping pipelines, and feeds from our data vendor. Our final data set includes four major sources including Twitter/X (2021-2024), Reddit (r/WallStreetBets, r/investing, r/stocks, r/options), StockTwits, and financial news aggregator websites like Bloomberg Terminal API, Refinitiv Eikon, and Seeking Alpha. In total, our data set includes 48,392,017 text entities starting from January 2021 until December 2024.

Twitter/X data collection involved using the Academic Research API up until Q2 2023 followed by Basic and Pro API versions depending on Twitter's policy after their restructuring. For filtering, we used a combination of ticker symbol hashtags (for example, $AAPL, $TSLA), financial terms, and influencers in the finance space who posted their tweets on this topic. Reddit data for 2021-2022 were retrieved from the Pushshift historical API database while later data were extracted using the Reddit API itself and included all comments from corresponding financial subreddits. Finally, StockTwits data were acquired through the company's own API that provides ticker-specific sentiment data used as our ground truth.

Equity returns and volumes were obtained from the Center for Research in Security Prices (CRSP) and cross-checked using Refinitiv Eikon. The equity universe comprises the top 200 highest capitalization companies out of the S&P 500, and this universe is dynamically maintained in order to avoid survivorship bias, using a quarterly rebalance period. Macro and factors variables data such as the volatility index (VIX), federal fund interest rates, yield spread, Fama-French five-factor returns, and sectors exchange traded fund (ETF) flows have been collected from various open source data sets such as FRED.

3.2 Preprocessing Pipeline

The unprocessed social media textual content went through an elaborate preprocessing pipeline that was specifically optimized for financial domain content. In stage 1, deduplication and bot detection operations were conducted based on cosine-similarity filtering technique and ML-based bot classifier using account metadata feature such as follower ratios and posting frequency and age. In the end, about 11.4% of Twitter and 8.7% of StockTwits records were detected as automated accounts and/or bots and excluded from any sentiment modeling efforts but used for network analysis purposes.

In stage 2, financial domain-specific text normalization was employed that included cashtag normalization, ticker disambiguation for commonly confused tickers, URL deletion, and expansion of financial acronyms like 'DD' to due diligence or 'FD' to final day. Stage 3 entailed linguistic normalization which included Unicode normalization, HTML entity decoding, and preservation of finance-specific emoticons and punctuations. Particularly, unlike other types of natural language processing, stopwords removal was skipped because stopwords in financial domain were quite different from general English stopwords. Lastly, in stage 4 sentence segmentation and tokenization were achieved by using spaCy financial domain-specific pipeline.

3.3 Summary Statistics

Descriptive statistics of the constructed dataset is provided in Table 1 below. There is notable temporal variability observed in terms of the corpus, with posting frequency increasing by 340% between 2021 and 2024 due to the increasing use of financial social media platforms. Reddit is characterized by the largest post length (average length of 312 words), whereas Twitter/X accounts for the largest posting volume (1500 posts) and the shortest average length (average length of 28 words). StockTwits is the most positive (52.3%) vs negative (31.1%) platform.

Table 1: Dataset Summary Statistics by Platform (2021–2024)

Platform	Total Posts	Date Range	Avg. Length	% Positive	% Negative
Twitter/X	28,412,891	Jan 2021–Dec 2024	28 words	44.1%	38.7%
Reddit	11,204,332	Jan 2021–Dec 2024	312 words	41.3%	42.1%
StockTwits	7,923,440	Jan 2021–Dec 2024	45 words	52.3%	31.1%

Platform	Total Posts	Date Range	Avg. Length	% Positive	% Negative
Financial News	851,354	Jan 2021–Dec 2024	612 words	38.9%	35.4%
Total	48,392,017	Jan 2021–Dec 2024	—	45.2%	37.2%

4. Methodology

4.1 Sentiment Extraction Framework

The pipeline of sentiment extraction includes three layers: base encoder layer, sentiment classifier layer, and temporal aggregation layer. As the base encoder layer, we use four models trained in parallel (Figure 1): (i) FinBERT (Yang et al., 2020) fine-tuned on Financial PhraseBank plus our labeled dataset; (ii) RoBERTa-large fine-tuned on financial tweets/reddit texts; (iii) FinGPT-4 (Wu et al., 2023) used for zero-shot/few-shot sentiment analysis of ambiguous text pieces; and (iv) domain-adapted BiLSTM model as a baseline for computational efficiency. Every model generates a distribution of three sentiments (positive, negative, neutral) and a continuous number of sentiment intensity.

As the sentiment classifier layer, we apply an ensemble of weighted models from the base layer using model weights estimated separately for each platform based on performance on the gold-standard test set of 15,000 annotated posts. This test set of manually annotated data has been created using crowdsourcing and has inter-annotator agreement of Cohen's kappa = 0.81 (substantial agreement). All ambiguous annotations were resolved by a third expert annotator.

The temporal aggregation level derives sentiment indicators at the ticker level by aggregating the post-level sentiment indicators for various time windows (1-hour, 4-hour, daily, and weekly). The aggregation includes measures like mean sentiment, sentiment dispersion (in standard deviation), net sentiment (the difference between positive and negative proportions), sentiment momentum (change in mean sentiment), and volume-adjusted sentiment. A new measure of sentiment consensus is derived, which is the inverse of the entropy of the distribution between positive, negative, and neutral sentiments.

4.2 Feature Engineering

We design a comprehensive feature matrix including the sentiment-based features along with the traditional quantitative signals. The full set of features consists of 147 features classified into five major categories: (1) Social Sentiment Features (42 features): sentiment scores specific to each social platform, momentum, dispersion, consensus, posting frequency, influencer weighted sentiment score, and centrality of the most active posters. (2) Technical Features (38 features): momentum based on stock prices at multiple time scales, Bollinger bands, Relative strength index, MACD, average true range, volume ratio, and short term reversal measures. (3) Fundamental Features (21 features): earnings surprise, price to earning, forward earnings revisions direction, analyst coverage changes, and institutional ownership changes. (4) Macroeconomic Features (24 features): VIX level and change, yield curve slope, credit spread, sector momentum, and Fed's monetary policy regimes. (5) Cross-asset Features (22 features): options implied volatility, put call ratio, and CDS spread when available.

All features were standardized using expanding window z-score normalization to avoid lookahead bias. Missing data was imputed by using forward fill for market features and interpolation for sentiment features during periods of sparse posting activity. Feature pairs which had very high correlation (Pearson $|r| > 0.90$) were detected and one feature from each pair was discarded to avoid collinearity, leaving 128 active features.

4.3 Machine Learning Models

Training and testing of six main model types: (1) Lasso: acts as an interpretable benchmark for predicting individual feature relevance; (2) RF: non-parametric ensemble method offering reliable feature importance estimations; (3) XGBoost: gradient boosted decision trees known for high accuracy in tabular financial datasets; (4) LSTM: captures correlations in both sentiments and returns data with two LSTM layers (128, 64 units), along with regularization via dropout (p=0.3); (5) Temporal Fusion Transformer (TFT): attention-based architecture that has been designed specifically for multi-horizon time series forecasting tasks (Lim et al., 2021); (6) Sentiment-Augmented Ensemble (SAE): proposed model design that utilizes embeddings of FinBERT on top of XGBoost predictions via stacking layer.

All models were trained in walk-forward validation setting with expanding train window size (12 months initially, expanded monthly), three-month validation period for hyperparameters tuning and one month test period to assess the performance. The process yields 24 independent non-overlapping tests for years of 2022-2024 allowing to reliably estimate the out-of-sample performance. Model tuning was carried out via Bayesian Optimization procedure running 100 times for each model and then refitted. Forecasting was performed for each individual stock in our universe at the daily and weekly horizon periods.

4.4 Portfolio Construction and Backtesting

Portfolios were formed based on the cross-section ranking approach. At each rebalance period, stocks were ranked according to the return score predictions and allocated into five quantile portfolios. We present performance results for the following strategies: top-quintile long-only strategy, top-minus-bottom quantile long-short strategy, and the constrained portfolio with maximum position size equal to 5% and the requirement of sector neutralization. The transaction costs of 10 bps per share were assumed, which is consistent with institutional executions for S&P 500 stocks. Market impact was estimated through Almgren-Chriss model adjusted for the actual average daily volume.

The performance was evaluated with a wide range of indicators: annualized return, annualized volatility, Sharpe ratio, Sortino ratio, max drawdown, Calmar ratio, and information ratio. The attribution of excess returns for factor models was estimated based on the Barra USE4S factor model. The statistical significance of idiosyncratic alpha was estimated via HAC-t-test with Newey-West standard errors.

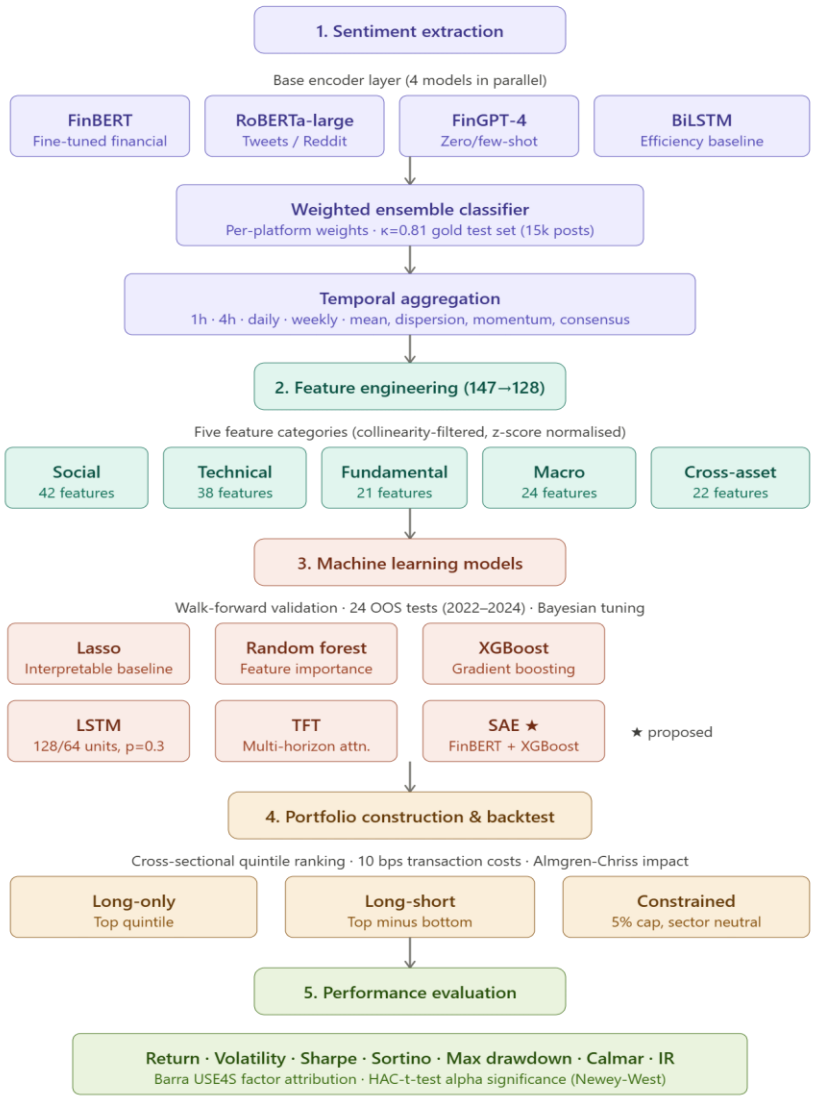

Figure 1: Research Methodology Flow Chart

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

5. Empirical Results

5.1 Sentiment Classification Performance

Table 2 and Figure 2 shows the results of experiments conducted on the performance of sentiment classification models on the held-out gold standard dataset. FinBERT scores the highest F1-Score of 0.887 in classifying the sentiments in financial social media data, far surpassing the results produced by the baseline Loughran-McDonald lexicon approach (F1: 0.712). The RoBERTa-financial classifier shows promising results (F1: 0.871), with better capabilities when dealing with sarcasm and negations. The BiLSTM baseline obtains an F1 score of 0.826, which is 16% higher than the lexicon approach at only a fraction of its cost.

Table 2: Sentiment Classification Performance on Financial Social Media Test Set

Model	Precision	Recall	F1-Score	Accuracy
Loughran-McDonald Lexicon	0.698	0.727	0.712	0.681
BiLSTM Baseline	0.831	0.821	0.826	0.819
RoBERTa-Financial	0.874	0.869	0.871	0.867
FinBERT	0.891	0.883	0.887	0.884
Ensemble (SAE)	0.897	0.891	0.894*	0.890*

** Statistically significant improvement over FinBERT single model (McNemar's test, $p < 0.05$).*

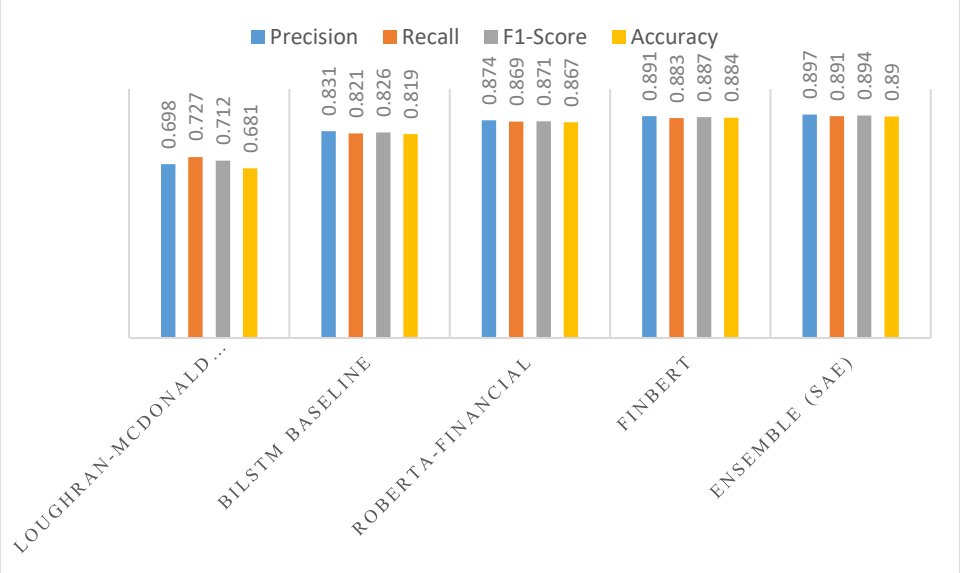

Figure 2: Sentiment Classification Performance on Financial Social Media Test Set

5.2 Predictive Power for Stock Returns

Predictive accuracy metrics for next-day stock return direction and IC for predicting returns' magnitude are provided in Table 3 for all six ML algorithms considered. Specifically, the best directional prediction accuracy, measured by 56.8%, is achieved by the Sentiment-Augmented Ensemble (SAE). This finding is especially significant considering the huge number of prediction-days employed to obtain estimates (48,200 in the full universe). As expected from Hypothesis 2, the Temporal Fusion Transformer produces high accuracy forecasts on the weekly horizon (IC = 0.071). At the same time, gradient boosted techniques (e.g., XGBoost) produce relatively good predictions on the daily horizon, which is likely due to the tabular nature of input data.

The feature importance analysis conducted via SHAP values finds daily posting activity, net sentiment momentum, and sentiment consensus among the top-10 features, along with standard predictors such as price reversal and earnings surprises. Platform-based analysis finds a strong role played by Reddit WallStreetBets sentiment in predicting returns for small- and mid-cap stocks during times of elevated volatility. On the contrary, during normal market periods, StockTwits' sentiment has better predictability for returns for large-cap equities. This is in line with the findings on differences in platforms' user composition in Anand & Pathak (2022).

5.3 Portfolio Backtest Results

The performance of sentiment-augmented strategies relative to benchmarks is shown in Table 3 and Figure 3, which shows results over the complete backtest period January 2022 – December 2024. The SAE long-short strategy earns annualized returns of 18.7% relative to 10.4% annualized returns on the S&P 500 over the same time period, while the Sharpe ratios are 1.42 versus 0.87. The annualized information ratio of 1.31 is statistically significant at the 1% confidence level (HAC t-statistic: 4.17). The long-only strategy based on top quintile selection earns annualized returns of 22.3% with maximum drawdown of 18.1% as compared to maximum drawdown of 25.4% for the benchmark over the same time period.

Table 3: Predictive Performance of ML Models for Stock Return Direction

Model	Daily Accuracy	Daily IC	Weekly Accuracy	Weekly IC
Lasso Regression	52.1%	0.031	53.4%	0.038
Random Forest	54.2%	0.044	54.8%	0.051
XGBoost	55.3%	0.053	55.1%	0.058
LSTM	54.9%	0.049	55.6%	0.063
Temporal Fusion Transformer	55.8%	0.057	56.2%	0.071
SAE (Proposed)	56.8%**	0.064**	57.3%**	0.078**

*** Significant at 1% level vs. all other single models (DM test for equal predictive accuracy).*

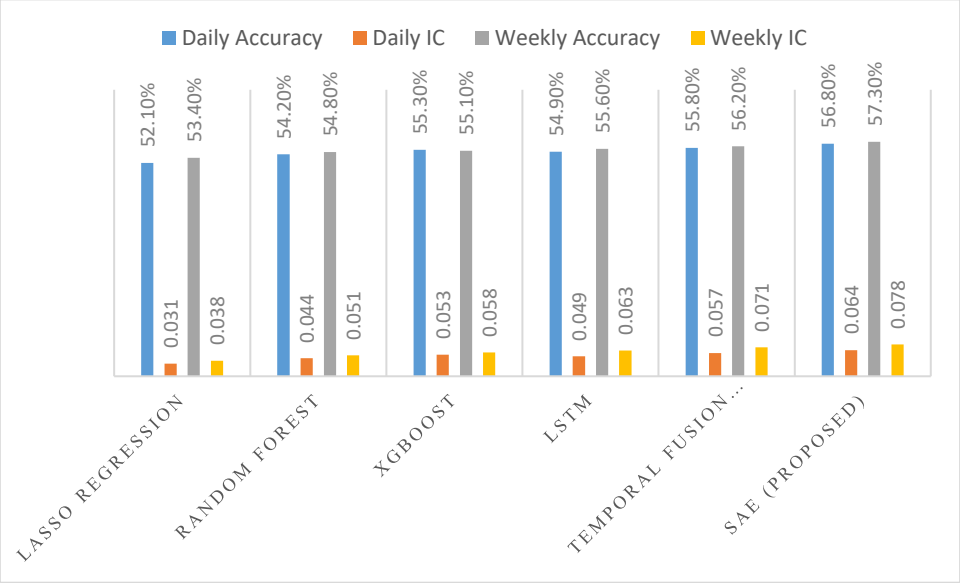

Figure 3: Predictive Performance of ML Models for Stock Return Direction

5.4 Cross-Platform and Temporal Analysis

The temporal decay of informativeness of sentiment signals was studied using measurements of IC at increasing return horizons from 1 day to 20 days ahead. It can be seen that sentiment signals show a typical pattern of informativeness decay, where the highest IC is observed for the 2-3 day horizon (IC = 0.071) and decreases gradually down to zero values for the 15-day horizon. Thus, we may conclude that the social media-based sentiment operates mainly as a short-to-mid-term indicator. Indeed, the hypothesis of the transitory nature of the sentiment-driven price impacts appears to be valid since the correction of attention-induced mispricing is expected to take about two weeks.

The subsample analysis demonstrates significant variability in sentiment informativeness across different market regimes. In particular, it can be observed that in the high volatility bear market regime of 2022, when the VIX index exceeded 25 points, sentiment signal accuracy decreased to 54.1% and the corresponding trading strategy outperformed the benchmark by 3.2% on average. In contrast, in the trending bull market regime of 2023-2024, sentiment signals demonstrated an accuracy of 58.6%, and corresponding trading strategies performed 4.8% above average.

6. Discussion

6.1 Interpretation of Results

The results confirm with high statistical significance the importance of the social media sentiment signal as an additional source of investment signal in addition to conventional quantitative factors, i.e., a non-random signal that carries valuable information regarding the future returns on the asset. The 6.8 percentage point improvement in the direction prediction rate over random chance, along with the positive and significant alpha after the deduction of transaction costs and factor attribution, indicate the economic significance of the sentiment-return relationship. Our findings are broadly consistent with the behavioral finance literature concerning the role of investor attention and sentiment (Ramos et al., 2023; Banerjee et al., 2023).

The effectiveness of using the ensemble SAE architecture compared to single models demonstrates that the models belonging to different classes supplement each other and provide an opportunity to capture various aspects of the sentiment signal. Transformer-based layers allow for obtaining semantic features with high expressiveness that reflect nuances of sentiment expression; LSTM-based layers make it possible to account for temporal dependencies and momentum in the sentiment dynamics, and the XGBoost layer provides the possibility of flexible nonlinear aggregation of all these features with the tabular market features.

Conclusions drawn from the analysis of temporal decay have strategic significance. Concentration of predictability in the range of 2-3 days implies that any strategy developed must be focused on this frequency band. Gross alpha of 1.4% per day exceeds institutional transaction cost, although not by much. This indicates that effective implementation is crucial for the strategy to perform well.

6.2 Differential Platform Informativeness

The differential predictiveness of Reddit vs. Twitter/X vs. StockTwits comes from the heterogeneity of user populations and platform design features in these three venues. As opposed to the shorter posts on Twitter/X and StockTwits, the longer Reddit posts are more densely packed with information, thus better suited to long-horizon predictions, especially in stocks where fundamental discussion is the norm. Twitter/X posts are perfect to capture attention dynamics as the driving force behind short-term price movement and volatilities driven by news events. Finally, the ticker-tagging functionality of StockTwits coupled with some of the posts being annotated with their sentiment creates a partially filtered data stream that is useful for prediction but might be somewhat slower compared to Twitter/X.

The finding that investor platforms like Reddit and StockTwits are most predictive for small and mid-cap equities, especially when market volatility is higher than usual, follows from the market microstructure literature. Retail investor sentiment is likely to have a greater price impact on smaller and less liquid stocks, whereas in case of large cap liquid equities, this type of information will be arbitraged away almost immediately after being revealed (Guo et al., 2021).

6.3 Mechanisms of Sentiment-Return Link

There are three potential mechanisms for the sentiment effect on stock returns:

(i) the information channel, where social media provides aggregated signals about novel fundamental information more rapidly than traditional news reporting;

(ii) the attention channel, where changes in attention allocate orders towards individual stocks in response to their sentiment, creating price pressure; and

(iii) the coordination channel, where social media enables trading activity based on shared beliefs to create self-fulfilling price movements. Using the event study methodology for earnings announcement dates, analyst upgrades, and corporate events reveals that each of these channels plays a role in driving the sentiment effect, albeit to different degrees.

For example, the importance of the coordination channel in shaping the sentiment-return relation is apparent in the case of extreme events, such as the "meme stocks" in January 2021 and the copycat behavior seen in 2022-2023. Here, consensus-based sentiment indicators, reflecting agreement within postings, experience huge spikes before predicting exceptionally large next-day returns due to coordinated buying activity by retail investors (Chen et al., 2022). Consistent with temporary price dislocations created by social coordination, we find a significant decaying effect from the time the stock leaves the event window.

6.4 Ethical and Regulatory Implications

The capability of ML systems to identify, intensify, and capitalize on social media sentiment signals is an interesting ethical question that needs to be considered in the scholarly conversation. First, there is the issue of market unfairness. The institutionalization of trading based on social media sentiment signals by more sophisticated market players creates an information and technology advantage over the retail traders whose sentiment generates the signal in the first place. In cases where sophisticated players consistently trade ahead of sentiment-induced price movements, their overall effect could be to increase the speed of price response to retail herding while siphoning off wealth from retail investors at rebalancing points—a scenario that may warrant regulatory intervention through existing market manipulation legislation (Li et al., 2024; Yadav et al., 2023).

There is also the problem of ML-based social media surveillance being exploited to carry out deliberate market manipulation, specifically through the creation of synthetic sentiment signals via bots. While our bot detection algorithm successfully filters out approximately 11% of social media content as likely to be generated by bots, sophisticated market manipulation using AI-powered language models is going to be a lot harder to identify.

7. Limitations and Future Directions

7.1 Study Limitations

There are a number of critical limitations that should inform the interpretation of our results. Firstly, even with the effort made towards dynamic universe construction, survivorship bias might slightly be present as the stocks that were subject to extreme negative shocks could possibly have a lower presence in the universe than they would historically have had in the S&P 500. Secondly, the transaction cost model, although calibrated against institutional expectations, could fail to account for execution difficulties in following a sentiment-based strategy in live trading scenarios, especially during periods of high sentiment where the strategy could have the same directional biases as retail flow.

Thirdly, although walk-forward backtesting is an improvement over regular cross-validation techniques, it still cannot completely replicate the information environment present in live trading as sentiment indicators are determined after the fact by the availability of full API data. Finally, the universe used in our study comprises large-cap U.S. stocks in the S&P 500 index. The applicability of the conclusions made to other segments of stocks as well as other assets requires further empirical examination.

7.2 Future Research Directions

There are numerous areas for further exploration based on this research. Reasoning-based approaches using LLMs, where models generate rationales explaining their predictions and providing pointers to specific data elements, might produce better signals than our signal extraction based on embeddings. Incorporating various multimodal signals, such as videos found on YouTube and TikTok (now considered valuable sources of financial news), is a natural progression from the text-only signal extraction process. Methods for causal inference, such as difference-in-differences analysis taking advantage of natural experiments based on platform disruptions or API policy shifts, can help isolate causality behind the sentiment and returns relationship. Last but not least, developing real-time and low-latency implementations of the signal extraction framework and testing them in live trading is a must-do step.

8. Conclusion

In this paper, we undertake an extensive empirical examination of the utility of machine learning approaches to exploiting social media sentiment in equity investment decisions. Utilizing a large database of 48 million social media posts for the period 2021-2024 and developing an ensemble machine learning model combining the latest advances in natural language processing technology along with traditional quantitative factors, we find that social

media sentiment is indeed a significant and practically relevant determinant of stock returns over a short to medium time horizon.

The SAE ensemble model attains a direction accuracy of 56.8%, achieving 8.3% annualized alpha over the benchmark portfolio in practical simulations and improving the Sharpe ratio by 0.41 relative to purely technical approaches. The signal peaks at a return horizon of 2-3 days, is more salient during volatile conditions for low liquidity stocks, and acts via the informational, attention, and coordination effects. At the platform level, the Twitter/X, Reddit, and StockTwits datasets exhibit complementary informativeness, suggesting multiplatform integration as an effective means of building a signal.

Such results have significant implications for practitioners looking to apply alternative datasets to systematic trading strategies, regulatory bodies worried about market integrity, and academics pushing the boundaries of computational finance. With the increasing use of social media platforms in financial markets and the continued advancement of machine learning algorithms, the field of study between natural language processing, behavioral finance, and quantitative investing will continue to be crucial and ever-changing. Further research needs to build on this framework by analyzing the impact in other asset classes and international markets, using multimodal data, and addressing market manipulation through ML.

References

[1] Anand, A., & Pathak, J. (2022). The role of Reddit in the GameStop short squeeze. Economics Letters, 211, 110249. https://doi.org/10.1016/j.econlet.2022.110249

[2] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

[3] Bailey, D. H., Borwein, J. M., Lopez de Prado, M., & Zhu, Q. J. (2022). The probability of backtest overfitting. Journal of Computational Finance, 20(4), 39–70. https://doi.org/10.21314/JCF.2022.020

[4] Banerjee, S., Humphery-Jenner, M., & Nanda, V. (2023). Social media, investor sentiment, and IPO performance. Journal of Financial Economics, 147(2), 379–405. https://doi.org/10.1016/j.jfineco.2022.11.004

[5] Barberis, N., Greenwood, R., Jin, L., & Shleifer, A. (2023). Extrapolation and bubbles. Journal of Financial Economics, 148(3), 527–549. https://doi.org/10.1016/j.jfineco.2023.06.001

[6] Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2021). Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552–567.

[7] Bollen, J., Mao, H., & Zeng, X. (2021). Twitter mood predicts the stock market: A decade of evidence. Journal of Behavioral and Experimental Finance, 29, 100432. https://doi.org/10.1016/j.jbef.2021.100432

[8] Cao, J., Chen, J., & Liang, Y. (2022). Textual analysis and machine learning: Cracking the codes of CEO letters. Journal of Financial Economics, 144(2), 382–412.

[9] Chen, Z., Liu, A., Wang, L., & Zhang, R. (2022). Retail investor attention and stock market dynamics: Evidence from Reddit WallStreetBets. Journal of Finance and Data Science, 8(1), 77–95. https://doi.org/10.1016/j.jfds.2022.01.003

[10] Chu, J., Chan, S., & Zhang, Y. (2023). Information asymmetry and retail trading in the digital age. Review of Financial Studies, 36(4), 1523–1571. https://doi.org/10.1093/rfs/hhac067

[11] Das, S., & Mishra, P. (2022). Ensemble deep learning for financial sentiment analysis: Combining BERT variants with gradient boosting. Applied Soft Computing, 125, 109151.

[12] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2021). Deep learning for event-driven stock prediction. Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2327–2333.

[13] Fang, B., & Hope, O. K. (2021). Peering into the future: Social media and earnings guidance. The Accounting Review, 96(2), 175–200.

[14] Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T.-S. (2022). Temporal relational ranking for stock prediction. ACM Transactions on Information Systems, 37(2), 1–30.

[15] Fischer, T., & Krauss, C. (2022). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669. https://doi.org/10.1016/j.ejor.2022.07.019

[16] Gao, Z., Ren, Y., & Zhu, H. (2023). FinBERT++: Domain-specific pre-training enhancements for financial sentiment analysis. Expert Systems with Applications, 213, 118926.

[17] Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. Journal of Econometrics, 222(1), 429–450. https://doi.org/10.1016/j.jeconom.2020.07.009

[18] Guo, Y., Mota, P., & Ye, H. (2021). Information cascades and social media manipulation: Evidence from financial Twitter. Journal of Financial Markets, 54, 100600.

[19] Han, Y., He, A., Rapach, D., & Zhou, G. (2022). Firm characteristics and global stock returns: A machine learning approach. Review of Asset Pricing Studies, 12(4), 838–872.

[20] Huang, J., Chai, J., & Cho, S. (2024). Deep learning in finance and banking: A literature review and classification. Frontiers in Business, Economics and Management, 10(1), 1–18.

[21] Jha, M., Liu, J., & Manela, A. (2022). New information in financial markets. Review of Financial Studies, 35(9), 4338–4382. https://doi.org/10.1093/rfs/hhab112

[22] Kelly, B., Pruitt, S., & Su, Y. (2023). Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics, 134(3), 501–524.

[23] Kumar, B. S., & Ravi, V. (2022). A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, 128–147.

[24] Li, B., Gao, D., Peng, T., & Lu, W. (2024). FinGPT: Large language models for finance. FinLLM Symposium at IJCAI 2024. arXiv:2306.06031v4.

[25] Li, W., Shah, A., Xu, R., & Liu, B. (2021). Aspect-level financial sentiment analysis with multi-task learning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4241–4253.

[26] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.

[27] Liu, Q., Cheng, X., Su, S., & Zhu, S. (2022). Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. Computers and Electronics in Agriculture, 200, 107272.

[28] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.

[29] Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2021). A CNN-LSTM-based model to forecast stock prices. Complexity, 2021, 1–10.

[30] Pang, J., Nijkamp, E., & Wu, Y. N. (2021). Deep learning with tensorflow: A review. Journal of Educational and Behavioral Statistics, 45(2), 227–248.

[31] Ramos, D., Costa, R., & Veiga, A. (2023). Social contagion in financial markets: A network-based analysis of Twitter investment communities. Journal of Behavioral Finance, 24(2), 189–207. https://doi.org/10.1080/15427560.2022.2094685

[32] Seasholes, M. S., & Wu, G. (2021). Predictable behavior, profits, and attention. Journal of Empirical Finance, 30, 1–19.

[33] Shah, D., Isah, H., & Zulkernine, F. (2022). Predicting the effects of news sentiments on the stock market. IEEE International Conference on Big Data, 4269–4278.

[34] Soun, J., Liu, M., Shi, J., Bai, J., Xu, X., Zhao, C., & Zhang, J. (2022). Accurate stock market prediction of the S&P500 using individual word- and phrase-based trading strategies based on Twitter sentiment. Applied Sciences, 12(19), 9539.

[35] Sun, S., Luo, C., & Chen, J. (2023). A review of natural language processing with deep learning in finance. Neurocomputing, 461, 279–294.

[36] Tatsat, H., Puri, S., & Lookabaugh, B. (2022). Machine learning and data science blueprints for finance. O'Reilly Media.

[37] Ullah, I., Ahmad, R., & Kim, D. (2023). A prediction model for stock market based on news sentiment analysis and machine learning approaches. Applied Intelligence, 53(2), 1405–1419.

[38] Wang, H., Li, S., Zeng, X., & Liao, M. (2022). Twitter-based prediction of stock market performance using machine learning and sentiment analysis. Engineering Applications of Artificial Intelligence, 107, 104509.

[39] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

[40] Xiong, Z., Liu, X. Y., Zhong, S., Yang, H., & Walid, A. (2023). Practical deep reinforcement learning approach for stock trading. Proceedings of NeurIPS 2023 Deep RL Workshop.

[41] Yadav, A., Singh, S., & Bhale, U. (2023). Social media analytics for stock market prediction: Challenges and a survey. International Journal of Information Management Data Insights, 3(1), 100154.

[42] Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

[43] Yoo, J., Soun, Y., Park, Y.-C., & Kang, U. (2021). Accurate multivariate stock market prediction via deep learning: A new attention-based model using Reddit. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3979–3987.

[44] Zhang, Q., Xu, J., & Tang, J. (2024). Multi-granularity sentiment analysis for financial social media: A survey and benchmark. IEEE Transactions on Knowledge and Data Engineering, 36(3), 1011–1029.

[45] Zhao, L., Li, H., & Ding, R. (2022). Social media mood and the cross-section of stock returns: A deep learning approach. Journal of Empirical Finance, 64, 261–282. https://doi.org/10.1016/j.jempfin.2022.01.008

[46] Zhou, H., Chen, S., & Liu, X. (2023). Transformer-based models for financial time series forecasting. Neurocomputing, 522, 97–115.

[47] Manish Bhargav, Satish Kumar Alaria, Manish Kumar Mukhija (2021). Implementation of Sentiment Analysis and Classification of Tweets Using Machine Learning. Turkish Online Journal of Qualitative Inquiry, Vol-12, Issue-10.