

News Sentiment, Social Media Public Opinion, and Stock Market Volatility: A Review of Multisource Data Forecasting Research

Zeichen Qu

Qingdao Academy, Qingdao, Qq052600, China

13173239517@163.com

ABSTRACT

The accelerating digitalization of financial information has positioned news sentiment and social media as increasingly influential forces shaping stock market dynamics. This review synthesizes recent interdisciplinary research that leverages multisource data—ranging from mainstream financial newswires to microblogging platforms and retail investor forums—to forecast market volatility. Unlike conventional surveys that concentrate narrowly on algorithmic optimization, we organize our discussion around three interconnected dimensions: the heterogeneous characteristics of sentiment signals across data sources, the theoretical mechanisms that translate public mood into trading behavior, and the practical challenges of integration in predictive modeling. Our analysis reveals that while individual data streams offer complementary predictive power, their fusion encounters persistent obstacles related to temporal misalignment, noise differentials, and interpretability trade-offs. We further identify that much of the existing literature remains preoccupied with short-term forecasting horizons, often neglecting the differential impact of sentiment on volatility regimes across bull and bear markets. This review concludes by outlining a research agenda that emphasizes causal identification strategies, multimodal sentiment decomposition, and adaptive modeling frameworks that can accommodate the rapidly evolving information ecology of modern financial markets.

Keywords: News Sentiment; Social Media; Stock Market Volatility; Multisource Data; Forecasting; Behavioral Finance

1. INTRODUCTION

For decades, financial economists have grappled with the fundamental question of what moves asset prices beyond fundamentals. The efficient market hypothesis, in its strong form, leaves little room for news or public chatter to exert meaningful influence, since all available information is presumably already discounted into prices. Yet anyone who has watched a stock tumble following an ambiguous tweet or surge on the back of a favorably framed earnings report knows that reality is messier than theory suggests. The growing accessibility of digital text data—from Reuters headlines to Reddit threads—has opened up new avenues for examining how language, emotion, and collective attention shape market behavior, and this line of inquiry has expanded rapidly over the past decade.

What makes this area particularly compelling is not merely the availability of data but the convergence of several intellectual currents. Behavioral finance has provided a steady stream of evidence that investors are subject to limited attention, confirmation bias, and herd behavior. Computational linguistics has furnished tools to quantify sentiment with increasing sophistication. And the sheer scale of social media activity has turned retail investors from price takers into agenda setters, as dramatically illustrated by the GameStop episode of early 2021. These developments have collectively pushed researchers to move beyond univariate sentiment indicators toward more ambitious frameworks that combine multiple data sources, each capturing a distinct facet of the information environment.

The present review attempts to provide some organizational clarity. Rather than offering an exhaustive catalog of every published model, we adopt a more selective approach that foregrounds conceptual issues we believe have been underappreciated. Specifically, we structure our discussion around the distinct characteristics of news versus social media sentiment, the theoretical pathways through which they influence volatility, and the practical difficulties entailed in combining them within a coherent forecasting framework. In doing so, we hope to strike a balance between accessibility for newcomers and critical engagement for specialists. We also deliberately foreground areas where the evidence is contested or where conventional wisdom may be overdue for revision. The review concludes with a set of concrete suggestions for future research that privilege causal inference over predictive horse races and contextual sensitivity over one-size-fits-all modeling ^[1].

2. RESULTS AND DISCUSSIONS

2.1 News Media: Institutional Voice and Deliberate Framing

Mainstream financial news has long served as a primary channel through which market-relevant information reaches investors. Unlike the cacophony of social media, news articles typically undergo editorial vetting, adhere to professional norms of attribution, and carry the institutional weight of established outlets. These features confer certain advantages for sentiment measurement: the language tends to be more measured, less prone to overt fabrication, and more readily interpretable through standard lexicons. Yet these same qualities also introduce systematic biases. News organizations are not passive conduits; they select which stories to cover, how prominently to position them, and which angles to emphasize. This framing effect can amplify certain types of information while suppressing others, and the resulting sentiment signal may reflect editorial posture as much as underlying economic conditions^[2].

Empirical studies have generally found that news sentiment possesses modest but statistically significant predictive power for short-term volatility, particularly around scheduled announcements such as earnings releases or macroeconomic data prints. The effect, however, appears highly asymmetric. Negative news tends to exert a stronger and more persistent impact than positive news, a pattern consistent with the well-documented negativity bias in human cognition. Moreover, the predictive relationship is not stable over time; it weakens during periods of high market uncertainty when investors become more reliant on other heuristics. There is also growing recognition that news sentiment should be disaggregated by topic rather than aggregated into a single polarity score, since the market's reaction to inflation news is qualitatively different from its reaction to geopolitical news, even if both carry a negative tone.

2.2 Social Media: Collective Emotion and Retail Voice

If news media represent the institutional voice of financial information, social media platforms embody its democratic—and frequently chaotic—counterpart. Platforms like Twitter, Reddit, and StockTwits host an incessant stream of opinions, rumors, jokes, and occasional insights, generated by a heterogeneous mix of amateur traders, professional investors, and automated bots. The sheer volume is staggering, but so is the noise. Sentiment extracted from social media tends to be more extreme, more transient, and more susceptible to manipulation than its news counterpart. Yet dismissing it as mere noise would be a mistake^[3]. There is mounting evidence that aggregated social media sentiment contains predictive information that is not fully captured by news wires, particularly for stocks with high retail ownership or those that are the subject of coordinated online campaigns.

One of the distinctive features of social media sentiment is its real-time nature. While news articles may take hours to be written, edited, and disseminated, a tweet can reach thousands of followers within seconds. This speed advantage makes social media particularly valuable for forecasting intraday volatility and for detecting emerging market narratives before they become mainstream. However, the temporal advantage comes with trade-offs. The lack of editorial oversight means that misinformation can spread rapidly, and sentiment spikes driven by non-informational factors—such as coordinated trolling or pump-and-dump schemes—can generate false signals. Researchers have responded by developing various filtering techniques, ranging from user credibility scoring to anomaly detection algorithms, though none has proven universally effective.

2.3 Comparative Dynamics and Complementarity

Rather than treating news and social media as competing alternatives, a growing number of studies have explored their complementary roles. The underlying logic is straightforward: news provides relatively reliable but lagging signals, while social media offers leading but noisy indicators. When combined judiciously, the two can offset each other's weaknesses. Several empirical investigations have indeed reported that hybrid models outperform those based on either data source alone, though the improvement is often modest and sensitive to the choice of aggregation window and sentiment extraction method. More interestingly, the nature of complementarity appears to vary across market conditions. During tranquil periods, news sentiment tends to dominate; during episodes of elevated uncertainty or when retail investor attention is unusually high, social media sentiment gains relative influence^[4].

There is, however, a notable gap in our understanding of how these two sources interact dynamically. Do social media sentiment spikes predict subsequent news coverage, effectively setting the agenda for mainstream media? Or does news sentiment shape the discussion on social media, providing a focal point for retail traders to react to? The evidence is mixed and likely bidirectional, but few studies have modeled this feedback loop explicitly. Similarly, the differential

impact of sentiment from various platforms—Twitter versus Reddit versus specialized investment forums—remains underexplored. These are not merely academic curiosities; they have practical implications for how forecasting systems should be designed and updated.

3. THEORETICAL FOUNDATIONS OF SENTIMENT-DRIVEN VOLATILITY

3.1 Behavioral Mechanisms Linking Emotion to Trading

The proposition that sentiment influences volatility rests on a chain of behavioral mechanisms that connect emotional states to trading decisions. At the individual level, investors in positive moods tend to exhibit overconfidence, underestimate risk, and trade more aggressively, which can amplify price movements in either direction. Conversely, negative sentiment fosters risk aversion, narrows attention, and triggers herding behavior as anxious investors seek safety in consensus. These individual-level tendencies aggregate into market-level phenomena through several channels. One prominent pathway is selective exposure: investors preferentially consume information that confirms their existing beliefs, and when sentiment becomes uniformly positive or negative, this selective exposure reinforces the prevailing mood and delays price discovery.

Another mechanism operates through shifts in the weight investors assign to different types of information. Under conditions of high uncertainty, investors rely more heavily on affective heuristics—essentially, going with their gut—and less on analytical reasoning. This makes sentiment a more powerful driver of volatility precisely when fundamentals are hardest to interpret. The implication is that sentiment and uncertainty are mutually reinforcing: uncertainty amplifies the influence of sentiment, and sentiment-driven volatility in turn heightens perceived uncertainty. This feedback loop can generate volatility clusters that persist well beyond the initial sentiment shock, a pattern that standard volatility models have difficulty explaining without resorting to ad hoc adjustments.

3.2 Attention Allocation and Information Processing

Before sentiment can influence prices, it must first capture investors' attention. This seemingly obvious point has profound implications. In an environment of information overload, attention is a scarce cognitive resource, and investors necessarily filter most available information. Sentiment signals that are more salient—whether due to extreme polarity, high volume, or vivid language—are more likely to break through this attentional bottleneck. This explains why sensationalist headlines often have outsized market impacts despite their limited informational content: they grab attention and trigger emotional responses before rational evaluation can occur^[5].

3.3 The Role of Noise Traders and Institutional Constraints

Any discussion of sentiment and volatility must grapple with the distinction between noise traders—who trade on sentiment rather than fundamentals—and institutional investors, who are subject to mandates, risk limits, and fiduciary duties. Noise traders are the primary conduit through which sentiment influences prices. Their trading activity introduces deviations from fundamental values, and when these deviations become large enough, they can trigger arbitrageur intervention. However, arbitrage is costly and risky, and in practice, constraints on short-selling and other institutional frictions prevent arbitrageurs from fully eliminating sentiment-driven mispricing. Thus, sentiment effects can persist, particularly in markets or sectors where arbitrage is most constrained.

4. METHODOLOGICAL APPROACHES AND PERSISTENT CHALLENGES

4.1 Sentiment Extraction: From Lexicons to Deep Learning

The operationalization of sentiment has evolved considerably over the past decade. Early studies relied on dictionary-based approaches, using financial lexicons such as Loughran-McDonald to map word frequencies onto polarity scores. These methods are transparent, computationally efficient, and easy to replicate, but they are brittle: they fail to capture sarcasm, negation, context-dependent meanings, or the subtle distinctions between different intensities of emotion^[6]. Subsequent work has adopted machine learning classifiers, trained on manually annotated corpora, which offer greater contextual sensitivity at the cost of requiring substantial labeled data and careful domain adaptation.

Most recently, transformer-based models—particularly variants of BERT fine-tuned on financial text—have achieved state-of-the-art performance in sentiment classification and have been increasingly adopted in volatility forecasting applications. These models are capable of capturing long-range dependencies and nuanced semantic relationships, but they are computationally intensive, require considerable expertise to implement, and are more difficult to interpret than

simpler alternatives. There is an unresolved tension between predictive accuracy and interpretability, and the choice of method often depends on whether the researcher's primary objective is out-of-sample forecasting or causal explanation. It is worth noting that many published comparisons between methods are not as conclusive as they appear, because performance differences are often sensitive to the choice of training data, preprocessing steps, and evaluation metrics.

4.2 Temporal Alignment and Frequency Mismatch

One of the most stubborn challenges in multisource forecasting is temporal alignment. News articles and social media posts arrive at different granularities—news at irregular intervals throughout the day, social media in a continuous, high-frequency stream—while volatility is typically measured at daily, hourly, or minutely intervals. Deciding how to aggregate sentiment over time introduces arbitrary choices that can materially affect forecasting outcomes. Should sentiment be accumulated with exponential decay to reflect fading relevance? Should overnight sentiment be treated differently from trading-hour sentiment? Should the timestamp reflect publication time or dissemination time? The literature offers little consensus, and many studies fail to justify their aggregation choices beyond appealing to convention^[7].

This temporal dimension is further complicated by the fact that sentiment effects are not instantaneous. Some news takes time to be absorbed and processed; social media sentiment may reach a critical mass only after several hours of discussion. The optimal lag structure likely varies by data source, market condition, and asset type. Yet most forecasting models impose fixed lags or rely on simple moving averages, which may wash out important timing effects. There is considerable room for methodological innovation here, particularly in the use of adaptive aggregation schemes or attention mechanisms that learn the relevant temporal dynamics from the data itself.

4.3 Evaluation Frameworks and the Overfitting Trap

A recurring concern in this literature is the risk of overfitting. With high-dimensional feature spaces, numerous model architectures, and ample opportunities for data mining, it is relatively easy to produce impressive in-sample results that evaporate in out-of-sample testing. Many studies report performance improvements that are statistically significant but economically negligible—a few basis points of forecast error reduction that would not justify the additional complexity in a real-world deployment. This issue is compounded by the use of non-standard evaluation metrics, incomparable test periods, and insufficient attention to transaction costs or market frictions.

More fundamentally, there is a lack of consensus on what constitutes a meaningful benchmark. Should a sentiment-augmented model be compared against a simple historical volatility forecast, a GARCH-type model, or a neural network trained on price data alone? Each choice yields a different baseline and can change the apparent value added of sentiment features. The field would benefit from more standardized evaluation protocols and a greater emphasis on replication studies. Until then, claims about the superiority of one method over another should be treated with appropriate caution.

5. CONCLUSION

The confluence of news sentiment, social media activity, and stock market volatility represents a vibrant and rapidly evolving research frontier. This review has argued that progress in this area depends on recognizing the distinct characteristics of different data sources, grounding empirical work in robust behavioral theory, and confronting methodological challenges with transparency and rigor. While the predictive potential of multisource sentiment data is real, it is also context-dependent, bounded by institutional constraints, and susceptible to overstatement. The most useful contributions in the coming years are likely to come not from ever-larger models or ever-more data, but from sharper conceptual frameworks, more careful empirical designs, and a renewed emphasis on the economic mechanisms that connect language to market outcomes. In an era of information abundance, the challenge is no longer finding data—it is making sense of it.

REFERENCES

- [1] Alomari M, Al rababa'a A R, El-Nader G, Alkhataybeh A. Who's behind the wheel? The role of social and media news in driving the stock–bond correlation. *Review of Quantitative Finance and Accounting*, 2021, 57(3): 959-1007.

- [2] Lehrer S F, Xie T, Zhang X. Social media sentiment, model uncertainty, and volatility forecasting. *Economic Modelling*, 2021, 102: 105556.
- [3] Deveikyte J, Geman H, Piccari C, Provetti A. A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 2022, 5: 836809.
- [4] Xu Y, Wang J, Chen Z, Liang C. Sentiment indices and stock returns: Evidence from China. *International Journal of Finance & Economics*, 2023, 28(1): 1063-1080.
- [5] Nuñez-Mora J A, Mendoza-Urdiales R A. Social sentiment and impact in US equity market: an automated approach. *Social Network Analysis and Mining*, 2023, 13: 111.
- [6] Lei B, Song Y. Volatility forecasting for stock market incorporating media reports, investors' sentiment, and attention based on MTGNN model. *Journal of Forecasting*, 2024, 43(5): 1706-1730.
- [7] Gupta T, Devji S, Tripathi A K. Investigating the impact of sentiments on stock market using digital proxies: Current trends, challenges, and future directions. *Expert Systems with Applications*, 2025, 285: 127864.