Isabella Lai

Sentiment Analysis is Virtually Useless in Financial Forecasting

The practical utility of sentiment analysis as applied in financial forecasting is limited, and often exaggerated in a potentially misleading way.

UNESCO / Dominique Roger, CC BY-SA 3.0 IGO <https://creativecommons.org/licenses/by-sa/3.0/igo/deed.en>, via Wikimedia Commons

Abstract

The practical utility of sentiment analysis as applied in financial forecasting is limited, and often exaggerated in a potentially misleading way. The costs of its implementation may often outweigh the potential benefits, and the absence of a causal link between financial news sentiment and stock market performance makes the observed correlation unsustainable. Through a detailed literature review, this paper presents an analysis of the common methodological flaws in existing studies, including data leakage, arbitrary lag selection, and inconsistencies in time frames. This paper concludes that sentiment analysis is not the ideal tool for facilitating financial forecasting, with an emphasis on the need for more transparent justification and methodological rigor, as well as the primacy of rational valuation within one's limited agency, over forecasting based on information sourced from layers of intermediaries.

Introduction

The role that sentiment analysis could possibly play in the context of forecasting invites scrutiny: while there are some correlations between news sentiment and market indexes (Antweiler and Frank, 2004), the real-world effectiveness appears constrained — the costs of implementation may often exceed the potential benefits, and public markets are not driven by individual sentiment (Kenyon-Dean *et al.*, 2018); news reflects the past; comments on the present through the lens of the past predicts the future when the news outlets have no agency in determining the new factors introduced as to what shapes the possible future.

The public market has evolved into an almost intelligent animal, with a mind of its own. Opportunities from mispricings persist, and the quest for the right price continues. Yet, the days when retail investors — who both feed into and follow the social media platforms, both the targeted consumer and ultimate content provider of the CNBCs — could, in practice, find such opportunities have long been receding. Over-sentimentalities can be augmented in concentrated crowds, but they dissipate before the news arrives. News of any kind is by its nature retrospective, under the guise of a comment of the present, and an unsolicited prediction of the future. Corpora of rhetoric, discontents, tarot readings, and others' wrongdoings, fitted curves almost invisible to the naked eye, after layers of fine-tuning, but the mind of the market, and the sentiment of an ultra-processed index, have never been a matter of vibes. It is tempting to ask why one would take them so seriously.

Despite, it makes sense intuitively to be related to investors' expectations; sentiment analysis is essentially a task of text classification (Kenyon-Dean *et al.*, 2018; Liu, 2020). Like the sentiment in the market, sentiment in NLP might not even be a metaphor, because a metaphor must be an alternative representation of something defined, or can in principle, be defined. Consensus exists about the nature and function of sentiments, but there is no consensual definition of what sentiments are. Similarly, the use of sentiment in the market has

This paper concludes that sentiment analysis is not the ideal tool for facilitating financial forecasting

been mostly a historical one, and today, not often used outside of motivated journalism.

The ambivalence toward sentiment analysis cannot be generalized to NLP itself, although sentiment analysis has been the most widely applied NLP task in finance. Natural Language Processing (NLP) does hold great potential here, but rather than appropriating metaphors, the prospect of natural language processing lies in the architecture of neural networks. As one example, a method to solve Initial Value Problems (IVPs), as related to PDEs using neural networks, may extend to pricing American options, especially with varying boundary conditions (Finzi *et al*., 2023). It's possible to create a hybrid model where the Neural IVP solver is used to handle certain continuous-time aspects of a problem, while a transformer deals with the sequential part. For instance, in a financial context, a transformer could be used to process discrete events (for instance, transactions) and analyze sequences of these discrete events to capture patterns and interdependencies with no intermediate metrics, while a Neural IVP solver could model continuous-time indicators, for instance, volatility surfacing, or interest rates, which are better modeled as continuous-time series.

The outperformance of transformers is partly due to the absence of a recognizable causal link between what goes in and what has been learned, a metric that reassures your intuitions. The neural network predicts neither a sentiment nor anything discrete. Regardless of how the architecture it relies on has evolved, sentiment analysis by itself is just a generic classification task for natural texts. It might have been a nice accessory in ancient Greek, but today, it is virtually useless in most of the tasks it promises to accomplish besides spam filtering. The idea of a sentiment lacks basis in its very existence. Like causality, there is no sentiment outside your own mental prison.

S&P as a miniature of the better parts of America

S&P 500 is often interpreted in two ways: a microcosm of the better-performing segments of the US economy, or as a benchmark against which mutual funds and companies measure performance, a practice that's standardized in SEC filings for listed companies. While commonly viewed as a passive market gauge, the S&P 500 actually reflects the outcomes of a specific, actively managed portfolio. Its composition is determined to a non-negligible degree by its committee, evolves over time, and tends to favor enduring companies with substantial connections.

The index's selection process provides a case for how the S&P 500 is massaged by the interplay between market capitalization weighting and the choice of the S&P index committee (S&P Global, 2023). Domicile considerations are not purely quantitative; the committee may weigh additional factors like operational headquarters and "market perceptions" in determining a company's eligibility (S&P Global, 2023). Such assessments might skew the index toward companies that, while possibly less representative of the broader market by standard metrics, align with the committee's holistic view of domicile relevance. With multiple share class lines, the committee can override standard market cap criteria, and consider liquidity and float-adjusted market cap instead, which can result in an index that favors certain trading volumes over a pure size representation, potentially giving more weight to companies with higher liquidity rather than larger overall market presence (S&P Global, 2023).

The ambivalence toward sentiment analysis cannot be generalized to NLP itself, although sentiment analysis has been the most widely applied NLP task in finance

Further, the discretion can be extended to adjustments in market capitalization ranges during the quarterly reviews. A deviation beyond a 10% threshold, which is not impossible, can prompt immediate recalibration of what constitutes an eligible company, shifting the market segment the index portrays (S&P Global, 2023). During mergers and acquisitions, the index committee can choose to include non-S&P 500 companies that acquire index constituents, regardless of whether all eligibility criteria are met, which could, under certain cases, lead to reduced turnover for maintaining market benchmark representativeness, but it does introduce an element of arbitrariness into the index's composition (S&P Global, 2023).

The S&P 500's reflection of the market cannot be reverse engineered from the outside; it is not a reflection but an actively shaped portfolio, occasionally commented on by the news, subject to policy interpretations and discretionary adjustments that can significantly influence its structural and performance characteristics. The news is a monologue commenting on an animal distant from its reach. It wouldn't be surprising if the distribution of the sentiment score of the news that contains 'sp500' as the keyword fits well with the distribution of its social media counterpart. But it would take substantially greater effort to fit it with price movement.

Literature review

Sentiment analysis as applied in financial forecasting is not a recent phenomenon. Earliest traditional NLP techniques faced challenges with long sentences and specialized text, whereas classical bag-of-words models were commonly used but had limitations. To overcome the shortcomings of traditional methods, recurrent deep learning models were introduced. The use of dictionaries and word embeddings, which consider linguistic similarities, enhanced the quality of models across various domains. Recent advancements have led to the development of transformer models that rely on an attention mechanism, and with consistently superior performance. FinBERT, a derivative of

BERT, was specifically pre-trained on financial text data for sentiment prediction.

Here, we specifically discuss only the top performance achieved by FinBERT and later replicate the methods in this study to find some insights (Araci, 2019). We selected this particular study for four main reasons: first, it represents the state of the art; second, it employs the most widely adopted model; third, it addresses a crucial task — identifying opportunities in the random walks through natural language processing; and fourth, while its merits are unique, its shortcomings are common to many studies that utilize sentiment analysis for financial forecasting.

Recent advancements have led to the development of transformer models that rely on an attention mechanism, and with consistently superior performance

In a 2022 paper, a team used sentiment scores extracted from financial news articles to predict the S&P 500 price direction with the state-of-the-art bidirectional encoder representations from transformers (BERT) models for sentiment classification (Devlin *et al.*, 2018; Fazlija and Harder, 2022). The researchers collected financial news data sets from Bloomberg and Reuters, including publication date, title, content, and sentiment classification. They also used the Financial Phrase Bank data set (Malo *et al.*, 2014). The study's initial focus was on the performance of the fine-tuned model for sentiment prediction of news items. The Financial Phrase Bank data set, with more than 50% agreement among annotators, was utilized for fine-tuning. The data set was divided into training, validation, and test sets in proportions of 72:8:20 (Fazlija and Harder, 2022). The target variable was the sentiment of the news, categorized into three labels: positive (label 0), negative (label 1), and neutral (label 2). The study also explored the prediction of the price direction of the next trading day using predicted sentiment scores with an implementation of a random forest. The models were first fitted in the initial four years and tested in the subsequent year, and in a second run, they were fitted in the first five years and tested in the sixth year, and so on. For classification tasks, the confusion matrix was used, containing information about true positives, true negatives, false positives, and false negatives, where a well-performing model would result in a confusion matrix with a high number of true positives and true negatives and a low number of entries off the diagonal (Fazlija and Harder, 2022).

The researchers further evaluated the performance of the FinBERT model, achieving an accuracy of 0.836, a weighted precision of 0.839, a weighted recall of 0.836, and a weighted F1-Score of 0.837. They compared various strategies based on sentiment scores, random forest classification on title sentiment scores, content sentiment scores, and a combination of both. From January 1, 2011, to August 16, 2016, the strategies based on content sentiment showed a return per annum of 15.208%, volatility of 15.097%, and a maximum drawdown of 20.598%. In contrast, the title sentiment alone resulted in a negative return per annum of -0.934%. When random forest classification was applied to title sentiment scores, the return per annum ranged from 7.673% to 9.920%, with corresponding volatility and drawdown figures. Strategies based on content sentiment scores yielded returns per annum of between 14.226% and 16.261%. Finally, when title and content sentiment scores were combined, the return per annum ranged from 13.093% to 14.143%. These quantitative findings illustrate the potential impact of different aspects of sentiment on stock price direction prediction, specifically, and quite unfortunately, in the context of the S&P 500 (Fazlija and Harder, 2022).

One minor issue lies in the use of k-fold cross-validation with time series data, which can lead to data leakage (Joseph, 2022). The 10-fold cross-validation used in this paper involves partitioning the data into 10 sets, or 'folds', and then running 10 separate training and validation cycles. In each cycle, a different fold is used as the validation set, while the remaining nine are used for training. The results are then averaged over the 10 cycles to produce the final metrics. However, if the data isn't shuffled or stratified before splitting into folds, some folds might be systematically different from others, leading to skewed training and validation. Further, the way data is distributed across folds matters (Joseph, 2022). For example, if all examples of a rare class end up in a single fold, the model might not learn to predict that class during most training cycles. If there's any overlap between the data used in different folds, or if the folds are not properly isolated, given the largely unstructured nature of the text data fetched directly from the news, it can lead to over-optimistic estimates of model performance. These concerns can have a direct impact on the reported metrics, and potentially render them not entirely accurate — there is no guarantee that the reported performance reflects the model's ability to generalize to future data.

The paper used the absolute volume of Bloomberg articles to determine information relevance, which may potentially misrepresent the days with significant news (Fazlija and Harder, 2022); the decision to standardize on 58 articles per day ignores the variability of news volume and could dilute important signals with noise; the random selection of articles when more than 200 are available could omit crucial information and add unnecessary variability.

Ignoring the fact that both sources are copyrighted, extending the Reuters data set beyond the Bloomberg data set's timeframe creates a temporal mismatch, compromising the synchronicity of the data. The Bloomberg data set ends on an earlier date than the Reuters data set. Extending the Reuters data beyond the last date of the Bloomberg data means that the additional data points from Reuters do not have corresponding data points in Bloomberg for comparison or joint analysis, and therefore, the model may learn features from the

Reuters data that are not present in the Bloomberg data set, creating an inconsistency in the feature space over time.

If the goal is to predict market movements based on news events, news from after the last Bloomberg date cannot be causally linked to market data prior to that date, because the market could not have reacted to news it had not encountered. Time series data have never been stationary, and the statistical properties change over time and financial time series would be simply ignored if they are to be stationary. A temporal mismatch means that the model's training data from one period (Bloomberg's time frame) may not represent the same underlying process as the data from another period (Reuters' extended time frame). The issue extends to assessing the model's performance, which can become challenging when the predictors (news articles) and targets (market responses) are from different time periods.

The ad hoc trial-and-error method for URL correction and the undefined "sufficiently large random sample" for content verification present significant methodological weaknesses (Fazlija and Harder, 2022). Ad hoc processes, by their nature, are not standardized and often not well-documented, making the experiment or data processing difficult to replicate, assuming yesterday's headline does deserve to be entertained today. The immediate issue is that without a clearly defined algorithm or rule set for URL correction, others cannot precisely recreate the data set, leading to potential inconsistencies in subsequent analyses. Verifying article correctness using an undefined sample size described as "sufficiently large" is methodologically unsound. Without a statistical rationale for the sample size, the chosen subset may not be representative of the entire data set, leading to selection bias. Key discrepancies or errors in the larger data set may go undetected if they are not present in the random sample, especially if the sample size is too small to capture the diversity of errors. Yet, none of them is addressed in the study — the absence of a replicable methodology for sample selection undermines the validity of any generalizations made about the data quality. The reliability of conclusions drawn from the sample about the entire data set depends on the sample being representative, which cannot be guaranteed without a proper sampling framework.

Regarding the inputs to the model, the authors have selected a data set for the model where sentiments agreed upon by just over 50% of annotators are deemed sufficient, despite available data sets with consensus rates of 100%, over 75%, and over 66% (Fazlija and Harder, 2022). This choice, assuming three distinct sentiment categories in news content, could be tenuous where annotator opinions are deeply split. Such a low threshold may not capture the true consensus, assuming it is possible on unstructured news headlines, potentially compromising the model's ability to accurately discern sentiment and introducing a bias towards the mildly majority-majority in cases of divided opinion (Geiger *et al.*, 2021). Although higher thresholds were considered, which would likely yield a clearer agreement and reduce ambiguity, the authors' preference for the >50% agreement data set suggests an optimization for favorable model performance metrics over a more representative data set. Absent a clear rationale provided in the paper, this decision could appear to prioritize appearance over scientific rigor, casting the model in a possibly misleading, favorable light.

If the goal is to predict market movements based on news events, news from after the last Bloomberg date cannot be causally linked to market data prior to that date

The selection of a uniform one-day time lag for predicting the price direction of the next trading day using predicted sentiment scores is a decision that raises some questions (Fazlija and Harder, 2022). The relevance of sentiment scores within such a time frame is ambiguous, and the relationships between news sentiment and stock price movements, if there are any reliable ones, are often event-dependent, varying across different time horizons. The study's lack of clear rationale for this specific time lag, coupled with the absence of robust validation and the exploration of trajectory-averaged displacements at varying lags, renders this approach seemingly arbitrary and counterintuitive. A standard justification for the lag selection would involve autocorrelation and cross-correlation analyses to determine the temporal alignment of sentiment scores and stock returns, alongside certain causality tests, although there aren't many, for predictive power assessment (Ritschel *et al.*, 2021). Rather than treating the biological clock of the news as objective, non-uniform, event-annotated study methodologies could provide empirical backing for the chosen lag by examining the stock's reaction to news events across different periods. Yet, without any underlying dependence between the predictor and the predicted, the model's efficacy is generically constrained.

Using yearly data to predict the next trading day's S&P 500 movement presents a methodological divergence from standard practices (Fazlija and Harder, 2022). Training on such extended time scales might not sensitize the model to short-term market volatility. A coarser temporal resolution than the prediction scale can induce spectral bias, where the model becomes attuned to slower-moving features, ignoring signals of higher frequencies that might capture some momentum effects. As a result, the model's parameters become skewed toward capturing long-term trends and autocorrelations, sacrificing the short-term dynamics. Further, without a sampling rate that captures the necessary signal frequency, predictions tend to be overly smoothed, masking critical market variations and undermining the model's performance in day-to-day "forecasting". Incorporating market microstructure noise might provide some help, but here, its omission from yearly data compounds

>

the risk of model misspecification by disregarding the intraday price dynamics and liquidity patterns. Yearly aggregates, although reflective of broader economic conditions, lack the immediacy of market sentiment and event-driven volatilities, making the theoretical basis for predicting daily S&P 500 fluctuations from such data questionable. With neither a coherent logic nor empirical substantiation, this approach risks bypassing every market signals that daily predictions are intended to exploit.

Finding significance by appropriating a metaphor from one discipline to the metaphor of another discipline isn't a counterargument to EMH

The last point of concern is the standard reading of the three levels of the efficient market hypothesis (EMH) presented at the beginning of their study. In one sentence, the source of the randomness in market behavior matters, whether it can be reduced to epistemic uncertainty or it is an intrinsic randomness. Random walk does not suggest that markets are entirely random, but rather that prices are in principle capable of fully reflecting all available information, and subsequent price changes can in principle represent all the random departures from previous prices (Malkiel, 2003). Market unpredictability stems from the arrival of new information, not intrinsic stochasticity. Under EMH, this unpredictability is not absolute; it's contingent on information dissemination rates and assimilation into prices. Prices then adjust to new equilibria, capturing information-based randomness (epistemic uncertainty as arises from incomplete knowledge) rather than ontological indeterminism (intrinsic randomness that implies an irreducible uncertainty inherent to market dynamics).

The standard assertion that markets are unpredictably random conflates the epistemic uncertainty inherent in price formation with the ontological randomness posited by intrinsic stochastic processes, overlooking that while market behavior is unpredictable to an extent, it is not indeterminate and is influenced by discernible information flows. Consequently, the challenge is not that markets can't be predicted, as the author of this paper interpreted, but that they can't be consistently outperformed without access to new information, which does not include the headline from yesterday, and subsequently prices cannot be predicted in a manner that yields consistent profits after considering risk and transaction costs (Malkiel, 2003). Finding significance by appropriating a metaphor from one discipline to the metaphor of another discipline isn't a counterargument to EMH. The market is an animal that does not speak your language.

To conclude with a relatively optimistic tone, the practical utility of sentiment analysis in financial forecasting remains contested, whether it is sourced from copyrighted news or social media corpus. Just like how a heavily processed index and a deeply massaged sentiment score might reassure you that the output is intelligible, and tomorrow be predictable, ignorance, especially the one uniformly distributed, can be equally reassuring.

REFERENCES

Antweiler, W. and Frank, M. Z. 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59(3). pp 1259–1294.

Araci, D. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv, abs/1908.10063.

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.

Fazlija, B. and Harder, P. 2022. Using Financial News Sentiment for Stock Price Direction Prediction. *Mathematics* 10(13). p. 2156. doi: 10.3390/math10132156.

Finzi, M., Potapczynski, A., Choptuik, M. and Wilson, A. G. 2023. A Stable and Scalable Method for Solving Initial Value PDEs with Neural Networks. arXiv preprint. Retrieved from https://arxiv.org/abs/2304.14994.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J. and Tang, R. 2021. Garbage in, garbage out revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies* 2(3). pp 795–827. doi: 10.1162/qss_a_00144.

Joseph, V. R. 2022. Optimal ratio for data splitting. Statistical Analysis and Data Mining. *The ASA Data Science Journal* 15. pp 531–538. doi: 10.1002/sam.11583.

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., Lalande, A., Bhanderi, S., Belfer, R., Kanagasabai, N., Sarrazingendron, R., Verma, R. and Ruths, D. 2018. Sentiment Analysis: It's Complicated! In Walker, M., Ji, H. and Stent, A. (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. *Human Language Technologies* Volume 1 (Long Papers). pp. 1886–1895. Association for Computational Linguistics. doi: 10.18653/v1/N18-1171.

Liu, B. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. In Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Studies in Natural Language Processing. Cambridge: Cambridge University Press. p. iii.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology.*

Malkiel, Burton G. 2003. The efficient market hypothesis and its critics. *Journal of Economic Perspectives* 17(1). pp 59–82.

Ritschel, S., Cherstvy, A. and Metzler, R. 2021. Universality of delay-time averages for financial time series: Analytical results, computer simulations, and analysis of historical stock-market prices. 10.1088/2632-072X/ac2220.

S&P Dow Jones Indices. 2023. Index Mathematics Methodology.