

QUANTIFYING THE VALUE OF FINE-TUNING: A REPRODUCIBLE FRAMEWORK FOR FINANCIAL SENTIMENT ANALYSIS IN THE TECHNOLOGY SECTOR

Marian-Pompiliu CRISTESCU^a, Dumitru Alexandru MARA^b, Ioana PETREA^c, Ana-Maria CONSTANTINESCU^{d*}

^{a)b)c)d)} "Lucian Blaga" University of Sibiu, Sibiu, Romania

Abstract

Financial sentiment analysis in the technology sector remains difficult because market-moving news combines firm-specific events, macro signals, and highly technical vocabulary. This study evaluates whether task-specific fine-tuning improves on zero-shot use of FinBERT when the corpus is restricted to English-language Technology news collected from MarketAux between 1 January and 30 June 2025. The final gold set contains 894 manually reviewed headline-description pairs labeled as negative, neutral, or positive and split into 625 training observations, 134 validation observations, and 135 test observations. On the held-out test set, zero-shot FinBERT reaches an accuracy of 0.637 and a macro F1 score of 0.626, whereas the fine-tuned version reaches 0.585 and 0.532. An exact McNemar test does not indicate a statistically significant difference between the two classifiers ($p = 0.450$). The findings suggest that, under a small sector-focused design and consumer-grade computational constraints, fine-tuning is not automatically superior to a strong domain-pretrained baseline.

Keywords: Sentiment analysis, FinBERT, domain adaptation, financial NLP, technology sector

JEL Classification: C55, G14, G17

* Corresponding author, Ana-Maria Constantinescu – anamaria.constantinescu@ulbsibiu.ro

1. Introduction

Financial news now reaches markets in real time and orders react within milliseconds. Manual reading cannot keep pace with this stream of headlines, social posts, and reports. Automated sentiment extraction from text has therefore become a core input to trading, risk management, and regulatory monitoring. Surveys and meta analyses show stable links between sentiment indices and returns, volatility, and trading volume across many markets (Du et al., 2024, Gric et al., 2023).

Early work on newspaper tone already showed that media pessimism predicts short term movements in stock indices and trading pressure (Tetlock, 2007). Later studies extended this line to survey based sentiment, online chatter, and intraday indicators. A recent meta analysis confirms that sentiment affects equity returns across many countries and measures, which supports the economic value of text based indicators (Gric et al., 2023). Newer research tracks real time public mood and finds that shifts in sentiment shape herding across investor types and industries, with strong effects in technology related sectors (Yang et al., 2025).

The technology sector sits at the center of these dynamics. It combines high valuation, dense media coverage, and highly specialized vocabulary. News on product launches, supply constraints, export controls, and artificial intelligence investment can alter expectations quickly even when the linguistic signal remains subtle.

Financial sentiment analysis has evolved from handcrafted dictionaries to large language models. Lexicon based tools remain popular thanks to their transparency and speed, yet they struggle with domain shift and context (Loughran & McDonald, 2011, Hutto & Gilbert, 2014). Transformer based encoders changed this picture. BERT and its descendants learn contextual token representations from vast corpora and support task specific fine tuning with modest labeled datasets (Vaswani et al., 2017, Devlin et al., 2019). FinBERT pushes this idea into finance by pre training on financial filings, news, and reports, and it improves over generic models on standard benchmarks such as Financial PhraseBank and FiQA (Araci, 2019).

Recent studies generally report competitive results for fine-tuned transformer models, yet the size of the gain varies with corpus size, label quality, and sector specificity. Evidence is less consistent when the supervised sample is relatively small or when the language mixes firm-level events with wider market signals.

This study therefore examines a restricted technology-news corpus and asks whether additional supervised adaptation yields a measurable advantage over direct zero-shot inference with a finance-pretrained encoder. The empirical setting is intentionally resource-aware and uses a reproducible workflow that can be executed on consumer hardware.

The analysis addresses three research questions. RQ1 asks how zero-shot FinBERT performs on manually labeled technology news. RQ2 asks whether fine-tuning on the same domain improves overall and class-level performance on a held-out test set. RQ3 asks whether any observed difference remains credible under an exact paired significance test. The study contributes a compact sector-focused benchmark, a transparent training configuration, and a reproducible comparison between zero-shot and fine-tuned FinBERT under modest computational constraints.

2. Literature review

2.1. Text based investor sentiment

Text based sentiment measures link language features to financial outcomes. Early studies used news counts or simple tone indicators and found that negative language in media columns predicts lower next day stock index returns and higher volatility (Tetlock, 2007). Later work broadened the evidence base by combining survey indicators, news, and search data into unified sentiment indices. A large meta analysis across 171 studies concludes that sentiment has a clear and robust effect on stock returns across markets, horizons, and data sources (Gric et al., 2023). Research on public mood and herding confirms that real time sentiment

shifts can amplify price co movement within sectors such as electronics and information technology (Yang et al., 2025).

The rise of big data and API based feeds extended available sources. Studies now use online news, earnings call transcripts, and social media content as parallel streams. Recent work builds daily or intraday indices from headlines and longer articles and shows that sentiment reacts strongly to volatility, crisis events, and policy surprises (Du et al., 2024, Cao, 2025). Cross market studies establish that sentiment measures derived from text can explain part of the variation in returns, volatility, and trading volume that remains after controlling for standard risk factors (Wu et al., 2025).

2.2. Lexicon based financial sentiment

Dictionary methods were the first systematic tools for financial sentiment analysis. General purpose resources such as Harvard IV often mislabel neutral financial terms as negative. The Loughran and McDonald dictionary corrected this issue by building word lists directly from 10 K and 10 Q filings, which improved performance in document level applications such as distress prediction (Loughran & McDonald, 2011). Tools such as VADER offer rule based sentiment scoring for short texts and remain widely used for social media and headline analysis thanks to their speed and open availability (Hutto & Gilbert, 2014).

Despite these strengths, lexicon methods face clear limits. They usually treat words as independent tokens and pay little attention to context, word order, or negation. They struggle with technical jargon and idioms that carry sector specific meaning. Surveys report that dictionary based approaches often underperform supervised methods on fine grained tasks, especially when texts contain subtle polarity shifts or mixed signals (Du et al., 2024). Extensions that blend finance specific lexicons with updated valence scores, such as FinVADER, improve results for some datasets yet still lag behind modern neural models for multi class classification (Du et al., 2024).

2.3. Machine learning and deep learning models

Statistical models such as logistic regression, support vector machines, and random forests formed the second wave of financial sentiment analysis. They rely on sparse features such as term frequency, TF IDF, or topic distributions. These methods already improved over pure lexicons on many tasks, yet their performance depends heavily on feature engineering and they still represent each word with a fixed vector that ignores context (Du et al., 2024).

Deep learning approaches introduced distributed word embeddings and sequence models such as convolutional networks and LSTMs. These architectures capture local context and non linear interactions and show strong results in sentiment classification and price prediction tasks. Studies that combine neural encoders for news with downstream models for returns or volatility forecasting report better predictive accuracy than shallow baselines, although gains vary with dataset size and label quality (Du et al., 2024, Cao, 2025).

2.4. Transformers, BERT, and FinBERT

Transformer encoders changed the landscape of natural language processing. The attention mechanism at the core of the Transformer allows each token to interact with every other token in the sequence, which yields better context modeling and faster training than recurrent networks (Vaswani et al., 2017). BERT introduced deep bidirectional pre training on large unlabeled corpora and a simple fine tuning scheme for downstream tasks (Devlin et al., 2019). BERT style models quickly set new benchmarks on many NLP tasks.

FinBERT applies this idea to finance. The original FinBERT work trained a BERT style encoder on a Reuters financial corpus and fine tuned it on Financial PhraseBank and other labeled datasets, achieving clear gains over both lexicons and generic BERT (Araci, 2019). Later variants trained on mixtures of filings, earnings calls, and analyst reports. These models offer stronger performance on document and sentence level financial sentiment tasks and feed into practical systems for news dashboards and risk monitoring (Araci, 2019, Du et al., 2024).

2.5. Fine tuning and domain adaptation in financial sentiment

The central question in current research is how much task specific fine tuning improves over zero shot use of pre trained models. Comparative studies in recent years address this for many datasets. A survey and benchmark by Du and co authors reviews lexicon, machine learning, and deep learning methods and concludes that fine tuned transformers consistently outperform static lexicons and shallow models across multiple languages and data sources (Du et al., 2024).

More focused experiments provide direct numbers. Jiang and colleagues use FinBERT with fine tuning to predict stock movements from news and show that adapted models reach higher F1 scores and directional accuracy than both lexicons and generic BERT based encoders (Jiang et al., 2023). Nasiopoulos et al. compare several fine tuned deep models, including transformer variants, on a suite of financial sentiment datasets and find that fine tuned encoders deliver the strongest macro average performance, especially for the minority class (Nasiopoulos et al., 2025).

Other work combines FinBERT with large language models. Shobayo et al. study stock index prediction using FinBERT, GPT 4, and logistic regression and report that, with careful hyperparameter tuning, a relatively simple classifier on top of well engineered features can rival more complex models (Shobayo et al., 2024). Shen and Zhang compare fine tuned FinBERT with zero shot and few shot large language models on news and reports and show that prompt engineered GPT 4o can approach the accuracy of a well tuned FinBERT model, yet at higher computational cost and with less predictable behavior across datasets (Shen & Zhang, 2024).

2.6. Domain and sector specificity

Many recent studies highlight the role of domain shift inside finance itself. Financial English for banking and macroeconomics differs from the language of high growth technology firms or derivatives markets. Work that analyzes Chinese financial news shows that language models tuned on local corpora perform better than generic multilingual models for both sentiment and price prediction (Chuang et al., 2025). Cross market studies that build federated models across exchanges stress the need to adapt representations to local language, sector structure, and news practices (Wu et al., 2025).

The semiconductor sector illustrates this problem clearly. Reports on chip shortages, fabrication capacity, export controls, and process technology use terms that often appear neutral to general models but signal clear good or bad news to domain experts. BBVA Research proposes an AI based framework that uses NLP and big data to track the global semiconductor sector and shows that news based indicators capture turning points in production and trade, especially during the recent supply shock (BBVA Research, 2023). Work on hype adjusted volatility measures uses news sentiment to separate temporary media excitement from information rich signals for technology stocks (Cao, 2025). These studies underline the need for sector aware sentiment tools.

2.7. Reproducible and resource aware financial NLP

Many state of the art transformer models require high end GPUs for fast training. At the same time, studies show that smaller models such as FinBERT, with roughly one hundred million parameters, can still be fine tuned on consumer grade hardware, sometimes with longer training times and careful memory management (Du et al., 2024). This trade off is central for researchers in smaller institutions and for practitioners who run models on on premise servers.

The literature indicates that financial text carries investable information and that transformer models benefit from domain pretraining. At the same time, the magnitude of any additional gain from task-specific fine-tuning remains contingent on corpus size, class balance, and sector-specific language. A compact, reproducible experiment on technology news is therefore useful for clarifying whether further supervised adaptation is warranted in a constrained applied setting.

3. Methodology

The empirical workflow follows a narrow sequence of data ingestion, corpus filtering, manual labeling, zero-shot evaluation, and supervised fine-tuning. The objective is to compare two FinBERT configurations on the same held-out technology-news split under a fully reproducible setup.

News records are collected through the MarketAux application programming interface. The final study corpus is restricted to English-language items classified in the Technology industry and published between 1 January 2025 and 30 June 2025. After filtering, deduplication, and manual review, the gold dataset contains 894 headline-description pairs labeled on a three-class scale: negative, neutral, and positive. The class distribution is 143 negative items, 469 neutral items, and 282 positive items.

Each observation is converted into a single model input by concatenating the headline and the short description. Manual annotation produces the final consensus label stored in the archived dataset. Individual annotator-level files were not retained in the reproducibility package, so a retrospective inter-annotator agreement coefficient cannot be reconstructed from the current materials and this remains a methodological limitation.

The dataset is split with stratification by class into training, validation, and test subsets, corresponding to 625, 134, and 135 observations. All runs use a fixed random seed of 42. Performance is evaluated on the same test split with accuracy, macro F1, per-class precision, recall, and F1, complemented by confusion matrices and an exact McNemar test for the paired comparison between zero-shot and fine-tuned predictions.

Table no. 1: Training configuration and evaluation split

Parameter	Value
Source model	ProsusAI/finbert
Industry filter	Technology
Language	En
Study period	2025-01-01 to 2025-06-30
Gold dataset size	894
Label distribution	143 negative; 469 neutral; 282 positive
Split	625 train; 134 validation; 135 test
Random seed	42
Maximum sequence length	160
Batch size	8
Learning rate	2e-5
Weight decay	0.01
Warm-up steps	50
Maximum epochs	8
Early stopping patience	2

Source: own research

The transformer backbone is ProsusAI/finbert. The fine-tuning configuration uses a maximum sequence length of 160 tokens, a batch size of 8, a learning rate of 2e-5, weight decay of 0.01, 50 warm-up steps, and a maximum of 8 epochs. Early stopping with a patience of 2 evaluation rounds is used to reduce unnecessary

training once validation performance stops improving. The workflow is designed to remain feasible on modest local hardware rather than specialized high-end infrastructure.

4. Results

The held-out comparison does not support a general performance advantage for supervised adaptation in this setting. Table no. 2 shows that zero-shot FinBERT attains an accuracy of 0.637 and a macro F1 score of 0.626, while the fine-tuned model reaches 0.585 and 0.532 on the same 135-item test split. The zero-shot configuration therefore delivers the stronger aggregate result.

Class-level evidence clarifies the pattern. Zero-shot FinBERT achieves very strong recall for the negative class at 0.909 and the highest overall balance across the three labels, whereas fine-tuning improves positive recall from 0.548 to 0.643 but materially weakens negative recall to 0.318. Much of the deterioration comes from negative and neutral items reassigned to the neutral or positive classes, as shown by the confusion matrices.

Table no. 2: Model performance comparison on the Technology test set

Model name	Accuracy	Macro F1 score
FinBERT zero shot	0.637	0.626
FinBERT fine tuned	0.585	0.532

Source: own research

Table no. 3: Per-class performance on the held-out Technology test set

Model	Class	Precision	Recall	F1	Support
FinBERT zero shot	negative	0.444	0.909	0.597	22
FinBERT zero shot	neutral	0.782	0.606	0.683	71
FinBERT zero shot	positive	0.657	0.548	0.597	42
FinBERT fine tuned	negative	0.412	0.318	0.359	22
FinBERT fine tuned	neutral	0.652	0.634	0.643	71
FinBERT fine tuned	positive	0.551	0.643	0.593	42

Source: own research

Table no. 4: Confusion matrices for the held-out Technology test set

Model	True class	Predicted negative	Predicted neutral	Predicted positive
FinBERT zero shot	negative	20	1	1
FinBERT zero shot	neutral	17	43	11
FinBERT zero shot	positive	8	11	23
FinBERT fine tuned	negative	7	14	1
FinBERT fine tuned	neutral	5	45	21
FinBERT fine tuned	positive	5	10	27

Source: own research

The paired exact McNemar test yields $p = 0.450$, with 28 cases corrected by fine-tuning and 35 cases worsened relative to zero-shot inference. The observed difference is therefore not statistically significant at conventional thresholds. Under the present data conditions, the fine-tuned model appears more specialized for positive language, but not more reliable overall.

The size of the aggregate difference is meaningful from an applied perspective, even though the paired test does not reject equality between the classifiers. Zero-shot FinBERT classifies 86 of the 135 test observations correctly, while the fine-tuned model classifies 79 observations correctly. Expressed as error rates, the zero-shot configuration misclassifies 49 items, compared with 56 errors for the adapted model. The absolute difference of seven correctly classified observations is modest in a small test set, yet it runs against the usual assumption that any task-specific training step should improve a domain-pretrained encoder. The macro F1 gap is wider than the accuracy gap because fine-tuning changes the distribution of errors across classes rather than merely shifting a small number of marginal cases. Accuracy is influenced by the dominant neutral class, whereas macro F1 assigns equal weight to negative, neutral, and positive items. The lower macro F1 of the fine-tuned model therefore indicates that the supervised update did not preserve class balance across the three sentiment categories.

The strongest contrast appears in the negative class. Zero-shot FinBERT identifies 20 of the 22 negative observations, which yields a recall of 0.909. The fine-tuned model identifies only 7 of the same 22 negative observations, producing a recall of 0.318. This shift is not offset by a comparable gain in precision. Negative precision falls from 0.444 under zero-shot inference to 0.412 after fine-tuning, so the adapted model becomes less capable of retrieving negative items while remaining only moderately precise when it predicts negativity. In financial news applications, this pattern matters because false negatives can suppress signals associated with deteriorating expectations, adverse regulatory developments, earnings pressure, or demand weakness. The zero-shot model is clearly more sensitive to adverse language in the present test split, although its sensitivity comes with a relatively high number of neutral and positive items incorrectly assigned to the negative class.

Neutral classification shows a different profile. The zero-shot model correctly assigns 43 of the 71 neutral items and produces a neutral recall of 0.606. The fine-tuned model correctly assigns 45 neutral items and raises neutral recall to 0.634, while neutral precision declines from 0.782 to 0.652. The fine-tuned classifier therefore labels slightly more true neutral observations correctly, but it does so by treating a larger share of non-neutral items as neutral. Fourteen of the 22 negative observations are moved into the neutral column after fine-tuning, compared with only one such reassignment under zero-shot inference. This indicates that the supervised update softens adverse signals rather than sharpening the separation between neutral and negative tone. For a technology-news corpus, where many headlines combine factual business updates with implicit valuation consequences, this movement toward neutrality suggests a conservative decision boundary learned from the small training set.

The positive class is the only area in which fine-tuning produces a visible recall gain. The adapted model correctly retrieves 27 of the 42 positive observations, compared with 23 under zero-shot inference, raising positive recall from 0.548 to 0.643. The improvement is narrow because positive precision falls from 0.657 to 0.551 and positive F1 remains almost unchanged, moving from 0.597 to 0.593. The confusion matrix explains this result. Fine-tuning sends 21 neutral observations to the positive class, while the zero-shot model sends 11 neutral observations to the same class. The adapted model therefore becomes more willing to predict positive sentiment, but part of this willingness comes from neutral items being interpreted as favorable. In a sector shaped by product announcements, investment narratives, and growth expectations, such a tendency may reflect recurring optimistic language in the training sample rather than a stable improvement in semantic discrimination.

The paired comparison provides a useful check on these descriptive patterns. Fine-tuning corrects 28 observations that the zero-shot model misclassified, yet it worsens 35 observations that the zero-shot model had classified correctly. The direction of this imbalance is consistent with the lower accuracy of the adapted model. The p value of 0.450 indicates that the imbalance is not large enough to support a statistically reliable difference in this test set. The result should therefore be read as evidence against a clear fine-tuning advantage rather than as definitive proof that zero-shot inference is universally superior. Within the present sample, the supervised update changes the model's behavior in class-specific ways, but the net movement is not strong or stable enough to justify replacing the zero-shot baseline on statistical grounds.

Taken together, the results point to a trade-off between sensitivity to negative language and broader adaptation to the observed training distribution. Zero-shot FinBERT appears to retain financial-domain priors that make it responsive to adverse expressions, even when those expressions occur in short technology-news items. Fine-tuning on 625 training observations shifts the model toward patterns visible

in the local corpus, especially positive and neutral phrasing. That shift may be useful when the operational objective prioritizes optimistic news detection, but it weakens the model for balanced three-class classification. The central empirical finding is therefore not simply that one model has a higher headline score. The more relevant result is that the supervised step redistributes errors in a way that reduces macro-level robustness under the available data conditions.

5. Discussion

The results qualify a common expectation in recent financial natural language processing research: additional supervised adaptation does not necessarily outperform a strong domain-pretrained baseline when the training corpus is narrow and imbalanced. The technology sample contains many mixed signals, where firm-specific news coexists with macroeconomic shocks, tariff developments, and broad market commentary. Under these conditions, the zero-shot model may preserve broader financial priors that remain useful on the held-out split, whereas fine-tuning can overfit recurrent local patterns.

This outcome does not diminish the practical value of the workflow. The pipeline remains lightweight, reproducible, and executable under modest hardware constraints, which matters for small research teams and applied analysts. Its main contribution lies less in claiming a universal gain from fine-tuning and more in showing when such a gain does not emerge, even after careful preprocessing and supervised adaptation.

Several limits should temper interpretation. The gold set contains 894 English-language items from a six-month window and a single industry filter, so external validity remains limited. Only consensus labels were preserved in the archived dataset, which means an inter-annotator agreement coefficient cannot now be recomputed. The study relies on one random split, so some variance may reflect sample composition rather than a stable ranking between the two models. Broader corpora, repeated or temporally ordered splits, and parameter-efficient adaptation strategies would provide a firmer basis for generalization.

The findings are consistent with a cautious reading of domain adaptation. FinBERT already encodes financial language through domain pretraining, so the zero-shot baseline is not a generic language model entering the task without prior exposure to market vocabulary. Its representations already capture many expressions associated with losses, downgrades, regulatory pressure, expansion, demand, investment, and earnings expectations. When the supervised corpus is small, the fine-tuning stage has limited room to learn genuinely new sector semantics. It can instead amplify recurrent lexical patterns that are specific to the training split. In the present experiment, this amplification appears most visible in the weaker treatment of negative news and the greater readiness to classify items as positive or neutral.

Class imbalance provides a plausible explanation for this behavior. The training corpus is dominated by neutral observations, followed by positive observations, while negative items form the smallest class. A model updated on such data can improve validation behavior by aligning with the more frequent labels, but that adjustment may reduce sensitivity to minority sentiment. The negative class is especially difficult because adverse financial tone is often compressed into short phrases, such as weaker guidance, delayed demand, investigation, margin pressure, regulatory scrutiny, or supply constraints. These phrases can be surrounded by otherwise neutral information about firms, products, or market context. If the training set contains relatively few negative examples, the fine-tuned model may learn a decision boundary that treats many adverse but concise signals as neutral background.

The technology sector intensifies this issue because its news language is frequently mixed. A single item may combine positive innovation signals with cost pressure, legal risk, or uncertainty about demand. Product launches and artificial intelligence investment announcements often carry favorable connotations, while the same articles may mention valuation concerns, chip shortages, export restrictions, or delayed revenue effects. A headline-description pair can therefore contain competing cues within a very short input. The zero-shot model may benefit from financial pretraining that keeps these cues more broadly separated. The fine-tuned model, by contrast, may attach excessive weight to locally repeated expressions of expansion, investment, or market optimism, especially when the labeled sample does not provide enough counterexamples for subtle negative cases.

The result should not be interpreted as a rejection of fine-tuning as a method. It indicates that fine-tuning has to be evaluated against a strong baseline and under the same data constraints faced by the intended application. Many studies report substantial gains from supervised adaptation when larger datasets, cleaner

labels, repeated validation schemes, or broader corpora are available. The present design differs because it uses a compact, sector-bounded dataset and a consumer-grade workflow. Under those constraints, the marginal benefit of updating all model weights may be lower than the risk of over-specialization. Parameter-efficient adaptation, class-balanced sampling, focal loss, threshold calibration, or label smoothing could produce a different balance between recall and precision, but those alternatives require separate testing rather than assumption.

For applied financial monitoring, the practical implication depends on the cost assigned to each error type. If the system is used as an early warning tool, missing negative items is usually more costly than over-flagging a small number of neutral items, because adverse news can require rapid human review. Under that objective, the zero-shot model is preferable in the present experiment because it retrieves most negative observations. If the system is used to detect positive momentum narratives, the fine-tuned model may appear attractive due to higher positive recall, yet the lower positive precision means analysts would face more false favorable signals. A single aggregate metric cannot settle this choice. The confusion matrix and class-level scores provide the more relevant basis for deployment decisions.

The absence of statistical significance deserves careful interpretation. The exact McNemar test evaluates paired disagreements and shows that the observed imbalance between corrected and worsened cases could occur by chance in a sample of this size. This does not erase the descriptive differences, because class-level shifts remain analytically important. It does prevent a strong claim that the zero-shot model is reliably superior across future technology-news samples. A larger test set or repeated splits could narrow uncertainty around the estimated gap. A temporally ordered evaluation would be especially valuable because financial language changes during earnings cycles, policy shocks, supply-chain disruptions, and sector-specific hype periods. Such a design would show whether fine-tuning improves adaptation to future news or merely memorizes patterns from the same time window.

The methodological limits are therefore directly connected to the empirical result. The absence of retained annotator-level files prevents a later assessment of disagreement around borderline items. This matters because financial sentiment labels are often contestable, especially when texts describe uncertainty, mixed earnings signals, or strategic investment with delayed payoff. A model may be penalized for predicting a defensible alternative label when the gold label itself reflects a difficult judgment. Preserving annotator-level decisions would allow future work to distinguish clear mistakes from ambiguous cases. It would support soft-label training, where uncertain items contribute probabilistic information instead of forcing a single hard class.

Several extensions follow from the present evidence. First, the negative class should be expanded through targeted sampling of adverse technology news, including regulatory actions, earnings misses, cyber incidents, supply bottlenecks, and demand revisions. Second, evaluation should report both random and temporal splits, since a model that performs well within one period may fail when sector narratives shift. Third, future experiments should compare full fine-tuning with lighter adaptation methods that reduce the risk of overwriting useful pretrained representations. Fourth, calibration analysis would clarify whether the models are merely changing labels or changing confidence in a systematic way. These extensions would preserve the reproducible character of the workflow while addressing the most important sources of uncertainty identified by the current results.

6. Conclusion

This study examined whether fine-tuning FinBERT on a manually labeled technology-news corpus improves sentiment classification relative to zero-shot inference. On the reproducible 894-item design used here, the answer is negative: zero-shot FinBERT attains higher overall accuracy and macro F1, while the paired difference between the two models is not statistically significant.

The result remains informative for financial natural language processing practice. A domain-pretrained model can remain competitive, or even preferable, when the available supervised corpus is small, imbalanced, and sector-bound. Future work should expand annotation breadth, preserve annotator-level labels, and test repeated or temporally ordered splits before stronger claims about sector-specific fine-tuning are advanced.

The empirical contribution lies in documenting the conditions under which fine-tuning does not create a measurable advantage. In the held-out comparison, the zero-shot model combines higher accuracy, higher macro F1, and substantially stronger negative recall. The fine-tuned model improves positive recall, but this comes with lower precision and a pronounced weakening of negative detection. The paired test confirms that the difference between classifiers is not statistically significant, so the evidence supports a restrained conclusion: supervised adaptation changed the distribution of predictions, yet it did not deliver a robust overall improvement for this technology-news corpus.

For researchers and practitioners, the main implication is that a domain-pretrained financial encoder should be treated as a serious baseline, not as a preliminary step that must automatically be replaced by a fine-tuned version. When labeled data are limited and class imbalance is visible, the baseline may preserve useful financial priors that a small supervised update can dilute. Before deployment, model choice should therefore be guided by class-level costs, paired testing, and inspection of error movement across sentiment categories. Future work can strengthen the evidence by increasing the annotated corpus, retaining annotator-level labels, testing temporally separated samples, and comparing full fine-tuning with parameter-efficient or class-balanced alternatives.

The conclusion drawn from this comparison is relevant precisely because it is modest. The experiment shows that additional training is not a purely technical formality, but a substantive modeling choice that can alter the sensitivity of a classifier in ways that matter for financial interpretation. In this case, the fine-tuned model became more responsive to positive wording while losing much of its ability to detect negative items. For technology news, this trade-off is consequential because adverse signals are often brief, embedded in otherwise factual reporting, and operationally important for risk screening. The evidence therefore supports a validation-first approach: a fine-tuned model should be adopted only after it improves the target metric and preserves acceptable performance on the classes with the highest practical cost. The present results favor zero-shot FinBERT as the stronger baseline for this dataset, while leaving open the possibility that larger, better balanced, temporally separated corpora could justify supervised adaptation. This distinction is important for reproducible financial NLP, where model complexity should follow demonstrated empirical value rather than methodological preference. Such evidence gives the study practical value for teams that must decide whether limited annotation resources should be invested in model adaptation or in broader validation.

References

- Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv*. <https://arxiv.org/abs/1908.10063>
- BBVA Research. (2023). *Tracking the semiconductor sector: An AI approach with NLP and Big Data*. Working paper.
- Cao, Z. (2025). *NLP forecasting of stock price volatility: A hype-adjusted probability measure and hype index* [Master's thesis, Birkbeck, University of London].
- Chuang, H.-M., He, H.-C., & Hu, M.-C. (2025). Chinese financial news analysis for sentiment and stock prediction: A comparative framework with language models. *Big Data and Cognitive Computing*, 9(10), Article 263. <https://doi.org/10.3390/bdcc9100263>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).
- Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9), Article 220. <https://doi.org/10.1145/3649451>
- Gric, Z., Bajzik, J., & Badura, O. (2023). Does sentiment affect stock returns? A meta-analysis across survey-based measures. *International Review of Financial Analysis*, 89, Article 102773. <https://doi.org/10.1016/j.irfa.2023.102773>
- Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 216–225).
- Jiang, Z., Xu, Y., & Zhang, H. (2023). Financial sentiment analysis using FinBERT with application in predicting stock movement. *arXiv*. <https://arxiv.org/abs/2306.02136>

- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Nasiopoulos, D. K., Roumeliotis, K. I., Sakas, D. P., Toudas, K., & Reklitis, P. (2025). Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models. *International Journal of Financial Studies*, 13(2), Article 75. <https://doi.org/10.3390/ijfs13020075>
- Ruan, L., Zhao, X., Liu, J., & Yu, J. (2025). Stock price prediction using FinBERT-enhanced sentiment with SHAP explainability and differential privacy. *Mathematics*, 13(17), Article 2747. <https://doi.org/10.3390/math13172747>
- Sharma, A. K., Ghosh, S., & Dey, L. (2025). Towards efficient FinBERT via quantization and coreset selection. In *Proceedings of the FinNLP Workshop*.
- Shen, Y., & Zhang, P. K. (2024). Financial sentiment analysis on news and reports using large language models and FinBERT. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)* (pp. 717–721). <https://doi.org/10.1109/ICPICS62053.2024.10796670>
- Shobayo, O., Adeyemi-Longe, S., Popoola, O., & Ogunleye, B. (2024). Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A data-driven approach. *Big Data and Cognitive Computing*, 8(11), Article 143. <https://doi.org/10.3390/bdcc8110143>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).
- Wu, L., Johnson, A. R., Carter, E. D., & Anderson, M. B. (2025). Cross-market financial sentiment tracking model based on federated learning and multimodal data. *Journal of Theory and Practice in Engineering and Technology*, 2(4), 1–8.
- Yang, Y., Lin, Y. W., & Cheng, L. C. (2025). Impact of real-time public sentiment on herding behavior in Taiwan's stock market: Insights across investor types and industries. *International Review of Economics and Finance*, 94, Article 104397. <https://doi.org/10.1016/j.iref.2025.104397>