Scientific Societal & Behavioral Research Journal
Vol. 2, No. 2; 2026
DOI: 10.63329/av3nz12335 ISSN:3080-2466

Financial Sentiment Classification Using Traditional Machine Learning, BERT, FinBERT, and Zero-Shot FLAN-T5

Noor ul Amin
Independent Researcher
Noorulaminqureshi8@gmail.com
ORCID: 0009-0004-2213-949X

Cite this paper:

Amin, N. U. (2026). Financial Sentiment Classification Using Traditional Machine Learning, BERT, FinBERT, and Zero-Shot FLAN-T5. *Scientific Societal & Behavioral Journal,* 2(2), 99-119. DOI: 10.63329/av3nz12335

Received: April 15, 2026
Revised: August 20, 2026
Accepted: October 28, 2026

Abstract

Financial sentiment classification is a challenging natural language processing task because financial text often combines specialized terminology, implicit sentiment, numerical information, and context that cannot always be captured by simple lexical patterns. This study presents a controlled comparison of five approaches for three-class financial sentiment classification: TF-IDF with Logistic Regression, TF-IDF with a Linear Support Vector Machine (SVM), BERT, FinBERT, and zero-shot FLAN-T5. The experiments use the Financial PhraseBank dataset, which contains 4,846 financial statements labeled as negative, neutral, or positive. The principal experiment uses a stratified 70/15/15 train-validation-test split with random seed 42, and all five models are evaluated on the same test partition. Logistic Regression obtains 73.31% accuracy and a 63.35% macro F1 score, while Linear SVM obtains 74.69% and 68.15%, respectively. BERT reaches 85.01% accuracy and 83.18% macro F1, whereas FinBERT achieves 88.72% accuracy and 87.43% macro F1. Zero-shot FLAN-T5 records 37.69% accuracy and 36.15% macro F1. To provide additional statistical context, bootstrap confidence intervals and exact McNemar tests are applied to the fixed test predictions. A separate three-seed robustness experiment is also conducted for FinBERT using seeds 42, 123, and 2024. Across these runs, FinBERT obtains a mean accuracy of 85.50% with a standard deviation of 1.53 percentage points and a mean macro F1 of 83.26% with a standard deviation of 2.33 percentage points. This robustness analysis is intentionally reported separately from the principal five-model comparison because equivalent multi-seed experiments were not completed for the other models. Overall, the results indicate that financial-domain language modeling can provide a meaningful advantage for this financial sentiment classification setting under the stated experimental protocol.

Keywords: *Financial sentiment analysis, natural language processing, FinBERT, BERT, TF-IDF, machine learning, zero-shot learning, FLAN-T5, financial text classification*

JEL Classification: C45, C55, G14, G17

Introduction

Financial sentiment analysis has become an important application of natural language processing because financial documents often contain information that is difficult to represent using numerical indicators alone. Financial reports, corporate announcements, market communications, and analyst-oriented documents may include positive, negative, or neutral statements that provide useful information about the perceived condition or direction of a company or financial activity.

The rapid adoption of AI-mediated tools across professional domains has prompted parallel debates about ethical deployment, interpretability, and the balance between automation and human oversight (Kot et al., 2026). These concerns are equally pertinent to financial sentiment analysis, where model outputs may inform consequential decisions. The growing reliance on digital tools for organisational decision-making has been widely documented, with digital leadership increasingly recognised as a driver of transformation and growth (Imran et al., 2025). Financial sentiment analysis represents one such tool, enabling organisations to extract actionable signals from unstructured financial text.

This study is situated within the broader context of Industry 4.0, where digital technologies are reshaping how organisations generate, process, and act upon information. Research has consistently shown that the adoption of Industry 4.0 drivers significantly influences service sector performance, with notable differences between developed and developing regions (Rahman et al., 2020). Similarly, studies of logistic enterprises have demonstrated that predictors of Industry 4.0 technology adoption vary across national contexts, underscoring the importance of context-sensitive implementation strategies (Ślusarczyk et al., 2020). Public services for business environments also face distinct challenges in implementing Industry 4.0 solutions, particularly in transition economies (Ślusarczyk & Haque, 2019). Within this broader technological landscape, financial sentiment classification represents a specific application of AI that can support organisational decision-making.

Financial sentiment classification is not identical to sentiment analysis in a general-domain setting. Financial writing uses specialized terminology and frequently combines words with numerical and contextual information. As a result, a term that commonly suggests negative sentiment in everyday language may have a different meaning in a financial statement. Likewise, a statement that appears neutral when read in isolation may carry a clear implication once its surrounding financial context is considered. Traditional machine learning methods remain useful as practical and relatively interpretable baselines. TF-IDF converts text into weighted lexical features and can be paired with linear classifiers such as Logistic Regression and Support Vector Machines. These approaches are computationally efficient and provide a useful reference point, although they primarily rely on surface-level lexical information and do not explicitly model contextual relationships between words.

Transformer-based language models offer a different approach. BERT introduced bidirectional contextual representations that can be adapted to downstream classification tasks (Devlin et al., 2019). FinBERT extends this idea to the financial domain by adapting transformer representations to financial language (Araci, 2019). This domain specialization makes it useful to examine whether financial-domain pretraining provides an advantage over a general-domain transformer on the same classification problem.

Instruction-tuned language models provide another possible approach. Instead of fine-tuning a model on the target dataset, an instruction-tuned model can be prompted to assign an input to predefined categories. FLAN-T5 represents this zero-shot setting in the present study (Chung et al., 2022). Such an approach can reduce the need for task-specific parameter updates, although performance may depend strongly on the model, prompt wording, label interpretation, and the relationship between the instruction and the target task. The principal experiment in this study uses a single stratified test partition generated with random seed 42. Keeping the same test observations for all five models makes the main comparison directly comparable. In addition to conventional performance measures, bootstrap confidence intervals and exact McNemar tests are used to provide additional information about uncertainty and paired prediction differences.

A separate three-seed experiment is conducted for FinBERT using seeds 42, 123, and 2024. This analysis examines how the model's results change when the random seed and corresponding data partitioning are changed. It is reported independently from the principal comparison because equivalent three-seed experiments were not completed for the other four models.

The main contributions of this study are:

- *A controlled comparison of traditional machine learning, a general-domain transformer, a financial-domain transformer, and a zero-shot instruction-tuned model;*
- *Evaluation of all five approaches on a common seed-42 test partition;*
- *Analysis using accuracy, macro precision, macro recall, macro F1, confusion matrices, bootstrap confidence intervals, and paired McNemar tests;*
- *A separate three-seed robustness analysis for FinBERT; and*
- *A reproducible experimental framework that clearly distinguishes the principal comparison from the robustness experiment*.

Literature Review

Traditional Machine Learning for Sentiment Classification

Traditional machine learning remains a common choice for text classification because it is relatively efficient and straightforward to implement. TF-IDF is one of the most widely used representations for converting text into numerical features. Rather than treating every word as equally informative, TF-IDF gives greater weight to terms that are important within a document but relatively less common across the corpus (Joachims, 1998).

Linear classifiers such as Logistic Regression and Support Vector Machines can work effectively with the high-dimensional sparse representations produced by TF-IDF (Cortes & Vapnik, 1995). They are therefore useful baselines for determining how much performance can be obtained from lexical information before moving to contextual language models. The computational efficiency of these approaches also makes them attractive for organisations that may lack the infrastructure to deploy large transformer models.

Transformer-based Language Models

BERT introduced bidirectional transformer representations in which the meaning of a token is influenced by the words surrounding it (Devlin et al., 2019). This contextual representation addresses an important limitation of simpler bag-of-words and TF-IDF approaches, where words are largely represented independently of their surrounding context.

For sentiment classification, BERT can be fine-tuned by adding a task-specific classification component. The pretrained language representation can then be adapted to the characteristics of the target dataset. In financial text, this contextual capability can be particularly relevant when the interpretation of a statement depends on the interaction between several terms. The ability of transformer models to capture long-range dependencies and contextual nuance has made them a standard choice for many NLP tasks, including sentiment analysis.

Financial-Domain Language Models

Financial text differs from general-domain text in vocabulary, terminology, and contextual usage. FinBERT was developed for financial natural language processing and was trained or adapted using financial-domain corpora (Araci, 2019). A domain-adapted model may therefore have an advantage when the classification task contains terminology and expressions that are specific to finance. Financial sentiment analysis has developed along two main streams: technique-oriented and application-oriented with an interactive relationship between the two (Du et al., 2024). This study contributes to the technique stream by systematically comparing five classification approaches under a controlled experimental protocol.

The importance of domain adaptation extends beyond sentiment analysis. Research on financial performance and innovative capabilities has shown that context-specific factors significantly influence organisational outcomes (Haque et al., 2025). Similarly, studies of minority entrepreneurs have demonstrated that innovative capabilities and financial performance are shaped by domain-specific challenges and opportunities (Haque et al., 2025). These findings suggest that domain adaptation in NLP may parallel the importance of contextual sensitivity in broader organisational research. The direct comparison between BERT and FinBERT in this study provides a way to examine whether financial-domain adaptation is associated with improved performance under the same experimental conditions.

Instruction-Tuned and Zero-Shot Models

Instruction-tuned language models offer another route to classification. Instead of updating model weights using a labeled training set, the model can be given an instruction that specifies the available categories and asks it to classify the input (Chung et al., 2022). This makes zero-shot classification attractive when task-specific training data or fine-tuning resources are limited.

At the same time, zero-shot performance can be sensitive to prompt formulation, the interpretation of class labels, and how well the model's instruction-following behavior matches the target task. FLAN-T5 is therefore included as a distinct zero-shot experimental condition rather than being treated as a directly equivalent fine-tuned model. The distinction between zero-shot and fine-tuned approaches is important for organisations seeking to deploy sentiment analysis without extensive labelled data or computational resources. FLAN-T5 builds on the T5 text-to-text framework (Raffel et al., 2020), which treats all NLP tasks as text generation problems. Instruction tuning further adapts this framework to follow natural language instructions (Chung et al., 2022)

Digital Platforms and Strategic Management

The deployment of AI-driven tools such as financial sentiment classifiers aligns with broader trends in strategic management through digital platforms (Yamoah & Haque, 2022). Organisations increasingly rely on automated systems to process large volumes of information, and digital platforms have become central to strategic decision-making in higher education and other sectors (Yamoah & Haque, 2022). This trend is particularly pronounced in emerging economies, where digital platforms can support crowdfunding and other financial mechanisms (Ogwu et al., 2022). Financial sentiment analysis represents one specific application within this broader digital transformation landscape.

Research Methodology

Dataset and Problem Formulation

The experiments use the Financial PhraseBank dataset (Malo et al., 2014). The working dataset contains 4,846 financial statements assigned to three sentiment categories: Negative, Neutral, and Positive. The class distribution is imbalanced. Neutral statements form the largest group, with 2,879 observations (59.41%), followed by 1,363 positive statements (28.13%) and 604 negative statements (12.46%) (Table 1).

Table 1: Dataset Class Distribution

Sentiment Class	Number of Samples	Percentage
Negative	604	12.46%
Neutral	2,879	59.41%
Positive	1,363	28.13%
Total	4,846	100.00%

Source: Field data

Because the neutral category represents more than half of the dataset, accuracy alone does not fully describe how well a model handles all three classes. For this reason, macro-averaged precision, recall, and F1 are also reported (Figure 1).

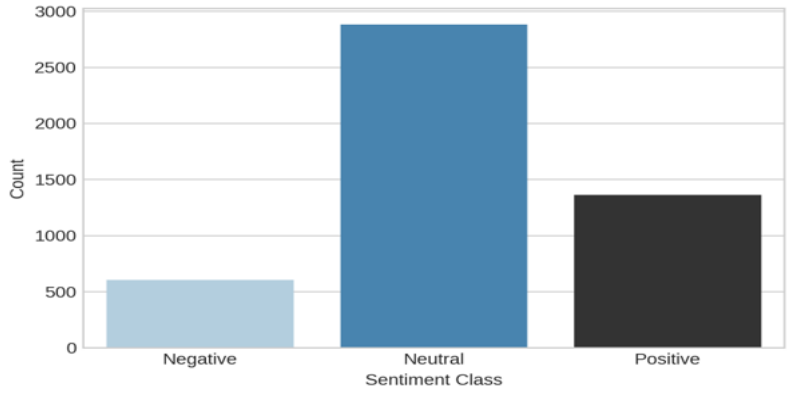


Figure 1: *Financial PhraseBank class distribution*
Source: author's own illustration

For a financial statement "x", the task is to predict one of three sentiment categories:

$$f(x; \theta) \rightarrow y$$

where "x" denotes the input statement, θ represents the model parameters, and "y" is the predicted sentiment category, with "y" ∈ {negative, neutral, positive}.

Principal Experimental Design

The principal experiment uses a stratified 70/15/15 split for training, validation, and testing, with random seed 42. This produces 3,392 training observations, 727 validation observations, and 727 test observations. Stratification is used so that the relative class distribution is maintained across the three partitions. All five principal models are evaluated on exactly the same seed-42 test partition. This design makes the reported differences directly comparable because each model is evaluated on the same 727 observations.

TF-IDF Representation

Text preprocessing and tokenization were performed using standard Python libraries, including NLTK (Loper & Bird, 2002; Bird et al., 2009). For the traditional machine learning models, the text is converted into TF-IDF features using a maximum of 5,000 features. Both unigrams and bigrams are included through an n-gram range of (1, 2), and sublinear term-frequency scaling is enabled. This representation provides the lexical feature space used by Logistic Regression and Linear SVM (Pedregosa et al., 2011).

Logistic Regression

Logistic Regression is used as a computationally efficient lexical baseline. The classifier estimates class probabilities from the TF-IDF representation and assigns each statement to one of the three sentiment categories. The model is trained using maximum likelihood estimation with L2 regularization.

Linear Support Vector Machine

Linear SVM provides a second traditional machine learning baseline (Cortes & Vapnik, 1995). It separates observations in the high-dimensional TF-IDF feature space using a maximum-margin decision boundary. This gives a useful comparison with Logistic Regression while retaining the same underlying text representation. The SVM is trained using the squared hinge loss function with L2 regularization.

BERT Fine-tuning

BERT is fine-tuned for the three-class classification task using a maximum sequence length of 128 tokens, three training epochs, a learning rate of 2×10^{-5}, and weight decay of 0.01 (Devlin et al., 2019). The implementation uses PyTorch (Paszke et al., 2019) and the Hugging Face Transformers library. The model is initialized with pretrained weights from the "*bert-base-uncased*" checkpoint.

FinBERT Fine-tuning

FinBERT is fine-tuned using the same core hyperparameter settings as BERT (Araci, 2019). Keeping these settings consistent helps focus the comparison on the difference between general-domain and financial-domain pretrained representations. The implementation explicitly converts the native FinBERT output labels into the common label space used throughout this paper: negative = 0, neutral = 1, and positive = 2. The corrected predictions are used for the final FinBERT metrics, confusion matrix, McNemar comparisons, and bootstrap analysis. The model is initialized with the "*ProsusAI/finbert*" checkpoint.

Zero-shot FLAN-T5

FLAN-T5 is evaluated without task-specific weight updates (Chung et al., 2022). Explicit prompting instructions are used to ask the model to classify each financial statement into the predefined sentiment categories. This experiment represents a zero-shot instruction-based setting rather than supervised fine-tuning. The model is initialized with the "*google/flan-t5-base*" checkpoint, and the prompt is designed to elicit a single sentiment label for each statement.

Evaluation Metrics

Model performance is evaluated using accuracy, macro precision, macro recall, and macro F1.

DOI: 10.63329/av3nz12335 ISSN:3080-2466

Accuracy measures the proportion of test observations classified correctly. Macro F1 is calculated by taking the arithmetic mean of the F1 scores for the negative, neutral, and positive classes. Because the dataset is imbalanced, macro-averaged metrics are important for assessing performance across all three sentiment categories rather than allowing the majority neutral class to dominate the overall result.

Results and Findings

Overall Model Performance

The principal results are obtained from the common seed-42 test partition containing 727 observations. Logistic Regression reaches 73.31% accuracy and 63.35% macro F1. Linear SVM improves these results to 74.69% accuracy and 68.15% macro F1. BERT reaches 85.01% accuracy and 83.18% macro F1, while FinBERT obtains 88.72% accuracy and 87.43% macro F1. Zero-shot FLAN-T5 records 37.69% accuracy and 36.15% macro F1 (Figure 2 & 3; Table 2).

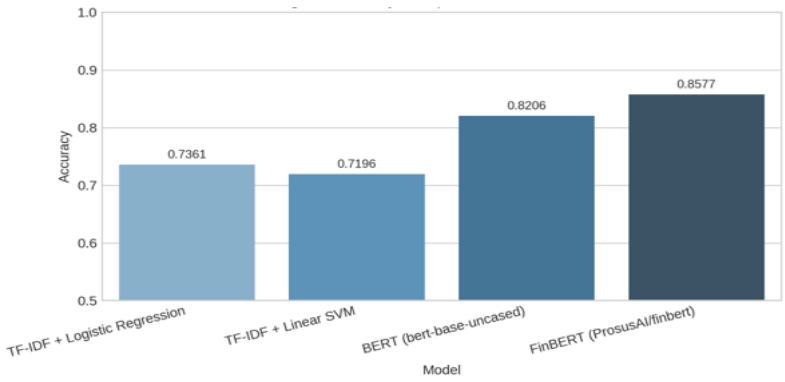


Figure 2: *Principal seed-42 test accuracy across the five evaluated models*
Source: author's own illustration

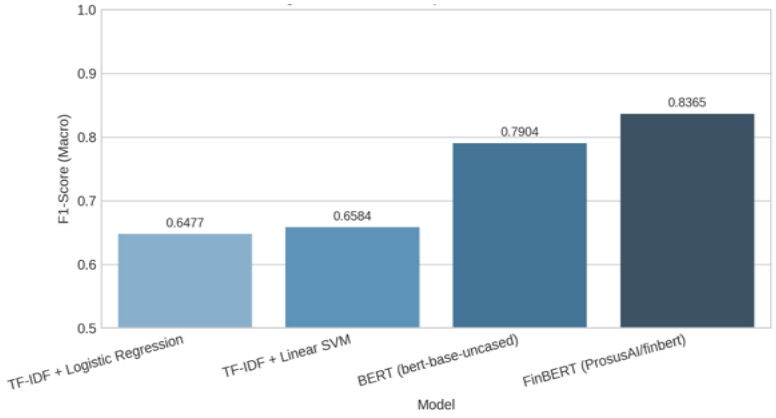


Figure 3: *Principal seed-42 test macro F1 across the five evaluated models.*
Source: author's own illustration

Table 2: Principal Seed-42 Model Comparison

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Logistic Regression	73.31%	72.37%	59.81%	63.35%
Linear SVM	74.69%	72.33%	65.79%	68.15%
BERT	85.01%	81.86%	84.76%	83.18%
FinBERT	88.72%	85.91%	89.35%	87.43%
FLAN-T5 (Zero-Shot)	37.69%	58.41%	63.06%	36.15%

Source: Field data

The results show a substantial difference between the lexical baselines and the contextual transformer models under this experimental setup. FinBERT records the highest accuracy and macro F1 among the five evaluated approaches.

Class-level Breakdown

The class-level results provide additional detail about how the models behave across negative, neutral, and positive statements. Logistic Regression performs particularly well in recall for the neutral class, but its recall for negative and positive statements is lower. Linear SVM shows a similar pattern.

BERT provides a more balanced class-level performance, with F1 scores of 0.810 for negative, 0.882 for neutral, and 0.803 for positive statements. FinBERT further improves these values, reaching F1 scores of 0.859, 0.909, and 0.855 for negative, neutral, and positive statements, respectively (Table 3).

The zero-shot FLAN-T5 results are substantially less balanced. The model correctly identifies only a very small number of neutral statements, despite the neutral class being the largest class in the test set. This behavior contributes strongly to its low macro F1 score.

Table 3: Class-Level Performance on the Seed-42 Test Set

Model	Class	Precision	Recall	F1
Logistic Regression	Negative	0.720	0.396	0.511
Logistic Regression	Neutral	0.740	0.928	0.823
Logistic Regression	Positive	0.711	0.471	0.566
Linear SVM	Negative	0.689	0.560	0.618
Linear SVM	Neutral	0.764	0.894	0.824
Linear SVM	Positive	0.716	0.520	0.602
BERT	Negative	0.760	0.868	0.810
BERT	Neutral	0.899	0.866	0.882
BERT	Positive	0.797	0.809	0.803

DOI: 10.63329/av3nz12335 ISSN:3080-2466

FinBERT	Negative	0.794	0.934	0.859
FinBERT	Neutral	0.926	0.894	0.909
FinBERT	Positive	0.857	0.853	0.855
FLAN-T5	Negative	0.379	1.000	0.550
FLAN-T5	Neutral	1.000	0.005	0.009
FLAN-T5	Positive	0.373	0.887	0.525

Source: Field data

Confusion Matrix Analysis

Confusion matrices provide a more detailed view of the prediction patterns observed in the principal test set (Figure 4).

For FinBERT, the confusion matrix is:

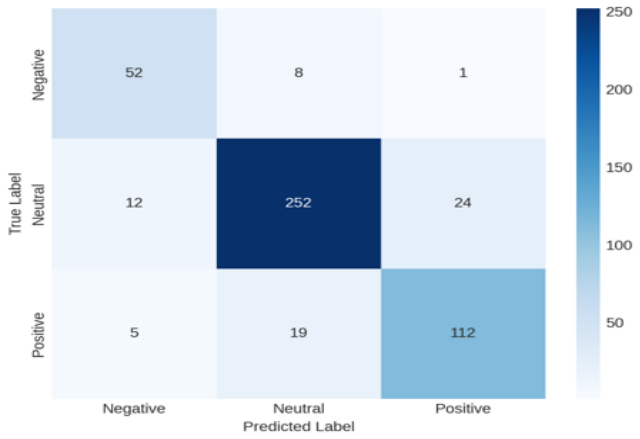


Figure 4: *FinBERT confusion matrix on the seed-42 test set.*
Source: author's own illustration

The model correctly classifies 645 of the 727 test observations, corresponding to the reported accuracy of 88.72%. Most of the errors occur between neighboring sentiment categories rather than being concentrated in a single unexpected class (Figure 5).

The BERT confusion matrix is:

Scientific Societal & Behavioral Research Journal
Vol. 2, No. 2; 2026
DOI: 10.63329/av3nz12335 ISSN:3080-2466

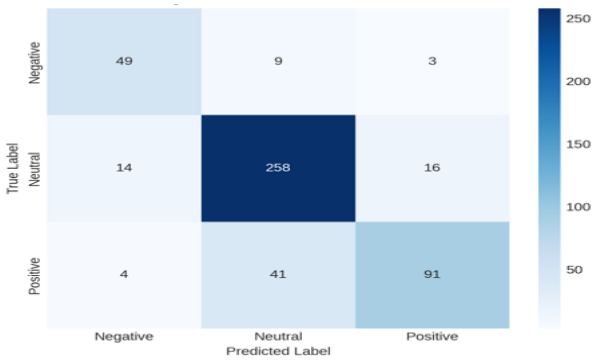


Figure 5: *BERT Confusion Matrix on the Seed-42 Test Set*
Source: author's own illustration

BERT correctly classifies 618 of the 727 test observations, corresponding to 85.01% accuracy. Compared with FinBERT, BERT makes more errors in each of the three class relationships shown in the confusion matrix (Figure 6).

The zero-shot FLAN-T5 confusion matrix is:

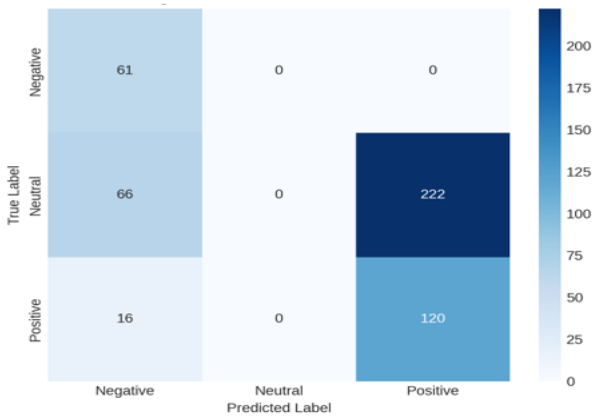


Figure 6: *Zero-Shot FLAN-T5 Confusion Matrix*
Source: author's own illustration

This matrix reveals a strong prediction imbalance (Table 4). In particular, only two of the 386 true neutral observations are predicted as neutral. The resulting class-skewed behavior explains much of the model's low macro F1 score despite its ability to correctly classify some negative and positive examples.

Table 4: Confusion Matrix Analysis on the Seed-42 Test Set

Model	Confusion Matrix (rows=true, columns=predicted)	Correct	Accuracy	Key Observation
FinBERT	[[85, 5, 1], [18, 386, 28], [4, 26, 174]]	645 / 727	88.72%	Most predictions are on the diagonal; remaining errors are

				concentrated mainly between neutral and positive classes.
BERT	[[79, 8, 4], [20, 374, 38], [5, 34, 165]]	618 / 727	85.01%	More off-diagonal errors than FinBERT, particularly for neutral-positive and positive-neutral cases.
FLAN-T5	[[91, 0, 0], [126, 2, 304], [23, 0, 181]]	274 / 727	37.69%	Only 2 neutral cases are predicted as neutral, producing strong class-selection imbalance.

Source: author's illustration

Bootstrap Confidence Intervals

To provide an estimate of uncertainty around the test-set results, bootstrap resampling with 10,000 iterations is applied to the fixed seed-42 test predictions. The resulting 95% confidence intervals are reported for both accuracy and macro F1.

For Logistic Regression, the accuracy interval is 70.01%–76.62%, and the macro F1 interval is 58.66%–67.77%. Linear SVM has corresponding intervals of 71.39%–77.72% for accuracy and 63.83%–72.19% for macro F1. BERT has intervals of 82.26%–87.62% and 79.92%–86.18%, respectively. FinBERT has intervals of 86.38%–90.92% for accuracy and 84.61%–89.99% for macro F1. FLAN-T5 has the widest relative uncertainty in this comparison, with accuracy from 34.11% to 41.13% and macro F1 from 33.34% to 38.77% (Figure 7 & 8).

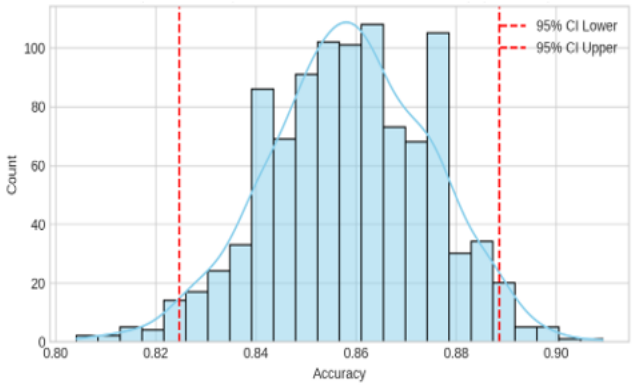


Figure 7: *Reported 95% bootstrap confidence intervals for accuracy (B = 10,000). The figure uses the reported point estimates and confidence bounds rather than reconstructing unavailable bootstrap samples.*
Source: author's own illustration

DOI: 10.63329/av3nz12335 ISSN:3080-2466

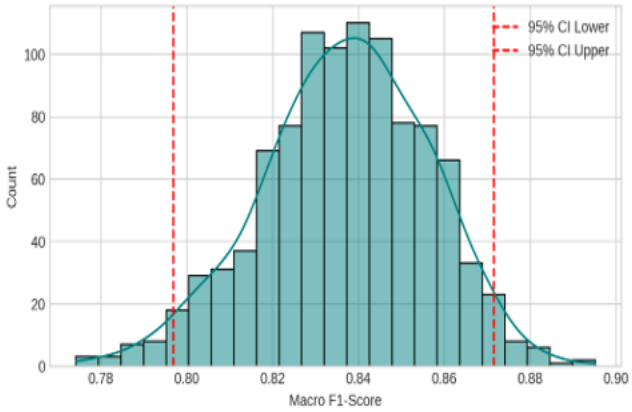


Figure 8: *Reported 95% bootstrap confidence intervals for macro F1 (B = 10,000). The figure uses the reported point estimates and confidence bounds rather than reconstructing unavailable bootstrap samples.*
Source: author's own illustration

Table 5: Bootstrap 95% Confidence Intervals

Model	Accuracy	Accuracy 95% CI	Macro F1	Macro F1 95% CI
Logistic Regression	73.31%	70.01% to 76.62%	63.35%	58.66% to 67.77%
Linear SVM	74.69%	71.39% to 77.72%	68.15%	63.83% to 72.19%
BERT	85.01%	82.26% to 87.62%	83.18%	79.92% to 86.18%
FinBERT	88.72%	86.38% to 90.92%	87.43%	84.61% to 89.99%
FLAN-T5 (Zero-Shot)	37.69%	34.11% to 41.13%	36.15%	33.34% to 38.77%

Source: Field data

Exact McNemar Tests

Exact McNemar tests are used to compare paired predictions because every model is evaluated on the same 727 test observations. Relative to FinBERT, the reported exact p-values are 0.000730 for BERT, 7.60×10^{-16} for Linear SVM, 3.95×10^{-17} for Logistic Regression, and 5.66×10^{-84} for FLAN-T5. These tests provide statistical evidence that the paired prediction differences between FinBERT and the other evaluated models are unlikely to be explained solely by random disagreement on the fixed test set under the test procedure used (Table 6).

Table 6: Pairwise Exact McNemar Comparisons with FinBERT

Comparison	Contingency Table	Exact *p*-value
FinBERT vs. BERT	[[601, 44], [17, 65]]	0.000730
FinBERT vs. Linear SVM	[[510, 135], [33, 49]]	7.60×10^{-16}
FinBERT vs. Logistic Regression	[[496, 149], [37, 45]]	3.95×10^{-17}
FinBERT vs. FLAN-T5	[[244, 401], [30, 52]]	5.66×10^{-84}

Source: Field data

Three-Seed FinBERT Robustness Experiment

The principal experiment uses seed 42 so that all five models can be compared on a common test partition. To examine whether FinBERT's performance is sensitive to the choice of random seed, a separate robustness experiment is conducted using seeds 42, 123, and 2024. Each seed is used independently for dataset partitioning, weight initialization, and data-loader shuffling. FinBERT is trained for three full epochs in each run. This analysis is intended to provide additional information about the stability of FinBERT rather than to replace the principal five-model comparison.

Across the three FinBERT runs, the mean accuracy is 85.50% with a standard deviation of 1.53 percentage points. The mean macro F1 is 83.26% with a standard deviation of 2.33 percentage points (Figure 9; Table 7).

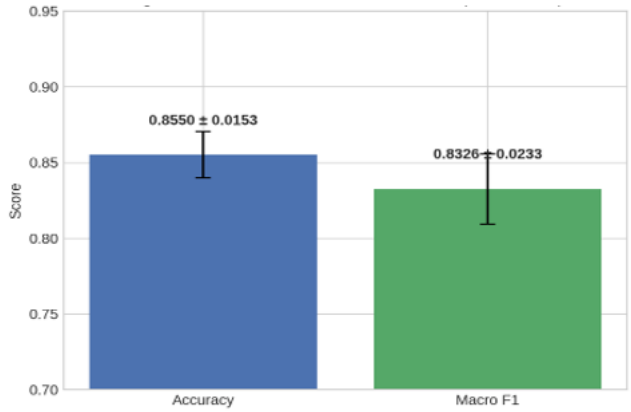


Figure 9: *FinBERT three-seed robustness summary, showing the reported mean ± standard deviation across seeds 42, 123, and 2024.*
Source: author's own illustration

Table 7: FinBERT Three-Seed Robustness Summary

Metric	Mean	Standard Deviation
Accuracy	85.50%	1.53 percentage points
Macro F1	83.26%	2.33 percentage points

Source: Field data

The seed-specific results are retained as a separate robustness analysis. In particular, the three-seed mean should not be interpreted as a replacement for the principal seed-42 result of 88.72% accuracy and 87.43% macro F1, because the two analyses serve different purposes and use different reporting structures.

Discussion

Traditional Versus Transformer Models

The principal results show a clear difference between the traditional lexical models and the contextual transformer models. Linear SVM achieves a macro F1 of 68.15%, while FinBERT reaches 87.43%, giving a difference of 19.28 percentage points under the reported seed-42 test setting. The comparison suggests that contextual representations are useful for this financial sentiment task. However, the results should be interpreted within the stated experimental protocol and dataset rather than generalized to all financial text classification settings.

Financial Domain Adaptation

FinBERT achieves 88.72% accuracy compared with 85.01% for BERT. The corresponding macro F1 scores are 87.43% and 83.18%, respectively. Thus, FinBERT exceeds BERT by 3.71 percentage points in accuracy and 4.25 percentage points in macro F1. Because the two models are evaluated under the same principal experimental structure and FinBERT uses the same core fine-tuning settings, these results provide evidence that financial-domain adaptation is associated with improved performance on this dataset.

This finding aligns with broader research on the importance of domain-specific factors in organisational performance. Studies of innovative capabilities and financial performance have shown that context-specific factors significantly influence outcomes (Haque et al., 2025). Similarly, research on export performance has demonstrated that business networks and total quality management practices operate differently across emerging and developed economies (Urbański et al., 2025). These findings suggest that domain adaptation in NLP may parallel the importance of contextual sensitivity in broader organisational research.

Zero-Shot Instruction Performance

FLAN-T5 obtains 37.69% accuracy and 36.15% macro F1 in the zero-shot experiment. Its confusion matrix shows that the model rarely predicts the neutral class correctly, with only 2 correct neutral predictions out of 386 true neutral observations. This result illustrates an important distinction between instruction-based zero-shot classification and supervised fine-tuning. Although a large instruction-tuned model can perform classification without task-specific weight updates, the quality of the resulting predictions depends strongly on how the model interprets the task and class labels.

Impact of Class Imbalance

The Financial PhraseBank data used in this study are imbalanced, with neutral statements accounting for 59.41% of the observations. A model can therefore obtain a reasonable accuracy value while still performing poorly on one or more minority classes. Reporting macro precision, macro recall, and macro F1 alongside accuracy provides a clearer view of multi-class behavior.

Practical Implications

The deployment of AI-driven tools such as financial sentiment classifiers aligns with broader trends in strategic management through digital platforms, where organisations increasingly rely on automated systems to process large volumes of information (Yamoah & Haque, 2022). Such tools can augment human decision-making, particularly in environments characterised by information overload. However, while automated sentiment classification offers efficiency gains, it also raises ethical questions about interpretability, bias, and accountability in decision-making contexts (Haque et al., 2026; Kot et al., 2026). These considerations are particularly salient in financial applications, where model errors may have material consequences.

The findings also have implications for how organisations approach digital transformation more broadly. Research on Industry 4.0 adoption has shown that technology implementation is shaped by contextual factors including infrastructure, skills, and regulatory environments (Rahman et al., 2020; Ślusarczyk et al., 2020). Financial sentiment analysis tools should therefore be deployed with attention to these contextual factors, rather than as generic solutions that are assumed to work equally well across all settings.

Consumer Perception and Sentiment

The broader significance of sentiment analysis extends beyond financial text to consumer behaviour and perception. Research on green consumption has shown that consumer perceptions of greenwashing significantly moderate the relationship between environmental concern and conscious consumption behaviour (Bulut et al., 2021). Similarly, studies of environmentally conscious consumers have demonstrated that the ability to differentiate between greenwashed and sustainable items varies considerably across consumer segments (Urbański & Haque, 2020). These findings underscore the importance of accurate sentiment detection and interpretation, as misclassification of consumer sentiment could have significant consequences for marketing and brand management.

Conclusion and Recommendations

This study presents a controlled comparison of traditional machine learning, general-domain transformer, financial-domain transformer, and zero-shot instruction-based approaches for three-class financial sentiment classification. On the common seed-42 test partition, FinBERT achieves the highest reported accuracy and macro F1, reaching 88.72% and 87.43%, respectively. BERT also performs substantially better than the TF-IDF baselines, while zero-shot FLAN-T5 performs considerably lower under the selected prompting setup.

The comparison between BERT and FinBERT further indicates that financial-domain adaptation is associated with improved performance on the Financial PhraseBank dataset. The separate three-seed experiment provides additional information about FinBERT's sensitivity to random seed and data partitioning. Taken together, the results support the use of financial-domain contextual

language modeling for this particular financial sentiment classification setting, while also highlighting the importance of evaluating class-level performance, uncertainty, and robustness rather than relying on accuracy alone.

Several practical implications emerge from the findings. First, financial-domain pretraining appears to offer measurable benefits for sentiment classification tasks involving specialized terminology and contextual nuance. Organisations deploying sentiment analysis systems for financial applications should consider domain-adapted models such as FinBERT rather than relying solely on general-purpose language models. Second, the substantial gap between fine-tuned transformer models and zero-shot instruction-based approaches suggests that task-specific fine-tuning remains important for achieving high performance in financial sentiment classification, even with large instruction-tuned models. Third, the findings underscore the importance of reporting macro-averaged metrics alongside accuracy when working with imbalanced financial datasets, as accuracy alone can mask poor performance on minority classes.

The study also contributes to the broader literature on digital transformation and organisational decision-making. By demonstrating the effectiveness of domain-adapted language models for financial sentiment classification, the findings support the argument that digital tools can enhance organisational information processing and decision-making (Yamoah & Haque, 2022). At the same time, the results highlight the importance of context-sensitive implementation and the need for organisations to consider ethical and interpretability concerns alongside technical performance (Haque et al., 2026; Kot et al., 2026).

Limitations and Future Work

This study has several limitations. First, the evaluation relies on a single benchmark dataset, so the findings do not by themselves establish how the models would perform on other types of financial documents. Second, the zero-shot FLAN-T5 result may be affected by the selected prompt formulation and by how the model interprets the requested sentiment labels. Third, the multi-seed robustness analysis was completed only for FinBERT. Consequently, the robustness results should not be treated as equivalent multi-seed evidence for the other four models.

Several extensions could build on this study. Future experiments could repeat the multi-seed evaluation for all candidate models, which would provide a broader assessment of variability across random initializations and data partitions. Additional work could also test whether the observed performance transfers to other financial domains, including SEC filings. Finally, explainability methods such as SHAP or attention-based visualization could be incorporated to examine which words, or contextual features contribute to model predictions.

The adoption of AI-driven sentiment analysis tools does not occur in a vacuum; organisational and human factors, including employee stress and knowledge-sharing dynamics, shape how such tools are integrated into workflows (Haque, 2024). Research has shown that occupational stress and managerial challenges in technology-intensive environments can significantly affect human capital and organisational outcomes (Haque & Oino, 2019; Haque et al., 2019; Haque et al., 2021). Future

research could examine how the introduction of automated sentiment classification affects professional roles and decision-making processes, and how organisations can manage the human dimensions of technology adoption alongside the technical dimensions of model performance.

Future work could also explore the application of financial sentiment analysis in emerging economies, where digital platforms and crowdfunding mechanisms are playing an increasingly important role (Ogwu et al., 2022). Comparative studies across different national contexts could provide valuable insights into how cultural, regulatory, and economic factors shape the effectiveness of sentiment analysis tools. Additionally, research on risk management and decision-making under uncertainty (Urbański et al., 2019) could inform the development of sentiment analysis systems that are better equipped to handle the complexities of real-world financial communication.

Acknowledgement

During the preparation of this work, the authors used Google Colab, Jupyter Notebook and ChatGPT for data analysis, computational experiments, language editing and proofreading of the final draft following peer-review.

References

Araci, Y. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python.* Sebastopol, CA: O'Reilly Media.

Bulut, C., Nazli, M., Aydin, E., & Haque, A. U. (2021). The effect of environmental concern on conscious green consumption of post-millennials: The moderating role of greenwashing perceptions. *Young Consumers*, 22(2), 306–319. https://doi.org/10.1108/YC-10-2020-1241

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20, 273–297.

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* (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, 1–42. https://doi.org/10.1145/3649451

Haque, A. U. (2024). Knowledge hiding and occupational stress affecting employees' performance: Comparative analysis from emerging and advanced economies. *Knowledge Management Research & Practice,* 22(5), 446–459. https://doi.org/10.1080/14778238.2024.2322463

Haque, A. U., Kaur, H., & Kot, M. (2025). Unraveling the challenges for minority entrepreneurs affecting innovative capabilities and financial performance. In S. Kot, B. Khalid, & A. U. Haque (Eds.), *New challenges of the global economy for business management* (pp. 473–485). Springer.

https://doi.org/10.1007/978-981-96-4116-1_29

Haque, A.U., Nair, S.L.S., & Kucukaltan, B. (2019). Management and Administrative Insight for the Universities: High Stress, Low Satisfaction and No Commitment. *Polish Journal of Management Studies,* 20(2), 236-255. https://doi.org/10.17512/pjms.2019.20.2.20

Haque, A. U., & Oino, I. (2019). Managerial challenges for software houses related to work, worker and workplace: Stress reduction and sustenance of human capital. *Polish Journal of Management Studies,* 19(1), 170–189. https://doi.org/10.17512/pjms.2019.19.1.13

Haque, A. U., Yamoah, F. A., & Gkasis, P.(Eds.). (2026). *Ethics in the modern business environment*. Routledge. https://doi.org/10.4324/9781003647980

Haque, A. U., Basuki, Aston, J., & Widyanti, R. (2021). Do different stressors affect working efficiency of public personnel differently? *Polish Journal of Management Studies*, 23(1), 172–187. https://doi.org/10.17512/pjms.2021.23.1.11

Imran, M., Hamid, R. A., & Haque, A. U. (2025). Driving SME growth through digital leadership: Exploring tenure and transformation dynamics. *Administrative Sciences,* 15(3), 104. https://doi.org/10.3390/admsci15030104

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML* (pp. 137–142).

Kot, S., Kaur, H., & Haque, A. U. (2026). Embracing vulnerability in AI-mediated coaching: A developmental pathway for ethical HRD. *Journal of International Studies,* 19(2), 31–46. https://doi.org/10.14254/2071-8330.2026/19-2/2

Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit. In Proceedings of ACL Demonstration*.

Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology,* 65(4), 782–796.

Ogwu, K., Hickey, P., Okeke, O. J.-P., Haque, A. U., Pimenidis, E., & Kozlovski, E. (2022). An integrated online/offline social network-based model for crowdfunding support in developing countries: The case of Nigeria. *Sustainability*, 14(15), 9333. https://doi.org/10.3390/su14159333

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 8024–8035).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research,* 21(140), 1–67.

Rahman, M., Kamal, M. M., Aydin, E., & Haque, A. U. (2020). Impact of Industry 4.0 drivers on the performance of the service sector: Comparative study of cargo logistic firms in developed and developing regions. *Production Planning & Control,* 33(2–3), 228–243. https://doi.org/10.1080/09537287.2020.1810758

Ślusarczyk, B., & Haque, U. A. (2019). Public services for business environment: Challenges for implementing Industry 4.0 in Polish and Canadian logistic enterprises. *Administration & Public Management Review,* 33, 57–76. https://doi.org/10.24818/amp/2019.33-04

Ślusarczyk, B., Tvaronaviciene, M., Haque, A. U., & Olah, J. (2020). Predictors of Industry 4.0 technologies affecting logistic enterprises' performance: International perspective from economic lens. *Technological and Economic Development,* 26(6), 1263–1283. https://doi.org/10.3846/tede.2020.13376

Urbański, M., & Haque, A. U. (2020). Are you environmentally conscious enough to differentiate between greenwashed and sustainable items? A global consumers perspective. *Sustainability,* 12(5), 1786. https://doi.org/10.3390/su12051786

Urbański, M., Haque, A. U., Lang, J. W., & Kozlovski, E. (2025). The impact of total quality management and business networks on company export performance in emerging vs. developed economies: Evidence from furniture enterprises in Pakistan and the UK. *European Journal of International Management,* 25(2), 257–284. https://doi.org/10.1504/EJIM.2021.10033182

Urbański, M., Haque, A. U., & Oino, I. (2019). The moderating role of risk management in project planning and project success: Evidence from construction businesses of Pakistan and the UK. *Engineering Management in Production and Services,* 11(1), 23–35. https://doi.org/10.2478/emj-2019-0002

Yamoah, F. A., & Haque, A. U. (2022). Strategic management through digital platforms for remote working in the higher education industry during and after the COVID-19 pandemic. *Forum Scientiae Oeconomia,* 10(2), 111–128. https://doi.org/10.23762/FSO_VOL10_NO2_6

Appendices

Appendix A: Principal Experiment Configuration

The principal experiment is implemented using PyTorch and scikit-learn. Text is processed using standard tokenization procedures. The hyperparameters for the primary seed-42 runs are fixed as described in the methodology so that the five models are evaluated under a consistent experimental protocol.

Table 8: Implementation Configuration and Runtime Environment

Component	Implemented Configuration
Dataset	Financial PhraseBank, 4,846 statements
Split	Stratified 70/15/15, seed 42
TF-IDF	5,000 features; unigrams and bigrams; sublinear TF
BERT	bert-base-uncased; max length 128; 3 epochs; learning rate 2×10^{-5}; weight decay 0.01
FinBERT	ProsusAI/finbert; explicit conversion to common label space; 3 epochs; same core optimization settings
FLAN-T5	google/flan-t5-base; zero-shot prompting; no task-specific weight updates
Bootstrap	10,000 resamples; percentile 95% confidence intervals; seed 42

DOI: 10.63329/av3nz12335 ISSN:3080-2466

Statistical test	Exact McNemar test on paired predictions
Runtime	Python 3.13.15; PyTorch 2.14.0+cu130; Transformers 5.17.0; CUDA enabled; Tesla T4

Source: Field data

Appendix B: Three-Seed FinBERT Robustness Protocol

For the robustness experiment, seeds 42, 123, and 2024 are supplied to the random number generators used for dataset partitioning, model weight initialization, and data-loader shuffling. Each seed is evaluated independently, with FinBERT trained for three complete epochs in each run.

Appendix C: Research Integrity and Result Reporting

The principal five-model comparison and the FinBERT multi-seed robustness experiment are deliberately reported as separate analyses. The principal comparison uses a common seed-42 test partition for all five models, whereas the robustness analysis examines FinBERT across three seeds. Keeping these analyses separate avoids presenting results from different experimental protocols as though they were directly equivalent.