Temporal Sentiment Shift Detection in Financial Markets: Domain Expertise vs Architectural Innovation

MSc Research Practicum
MSc Fintech Information

Pallavi Guduguntla
Student ID: x23357673

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	… Pallavi Guduguntla ………………………………………………………………………
Student ID:	…… x23357673………………………………………………………………………..……
Programme:	……… MSc Fintech Information……………………… **Year:** ……2024-25…..
Module:	… MSc Research Practicum ………………………………………….………
Supervisor:	… Brian Bryne ……………………………………………………………….………
Submission Due Date:	……11-08-25…………………………………………………………………….………
Project Title:	Temporal Sentiment Shift Detection in Financial Markets: Domain Expertise vs Architectural Innovation
Word Count:	……6902……………… **Page Count**……23…………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pallavi Guduguntla ……………………………………………………………………………………………

Date: 11-08-2025 ……………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Temporal Sentiment Shift Detection in Financial Markets: Domain Expertise vs Architectural Innovation

Pallavi Guduguntla
Student ID- x23357673

Abstract

This research investigates relative effectiveness of domain-specific pre-training against architectural innovations in detecting temporal shifts of financial sentiment across various regimes of markets. While models constructed using transformers have revolutionized analysis of financial sentiment, their ability of detecting shifts of sentiment during regime shifts of markets is not deep enough with related research. This study addresses this gap by systematically comparing FinBERT's financial domain expertise against ModernBERT's architectural advances including extended context windows, Rotary Positional Embeddings, and alternating attention mechanisms. Using a comprehensive dataset of 37,509 financial news articles from 2009-2023, augmented with market regime classifications and temporal annotations, three models were fine-tuned and evaluated: FinBERT, ModernBERT-base, and ModernBERT-large. The evaluation framework incorporated traditional classification metrics, specialized temporal shift detection measures, and cross-regime performance analysis. Results demonstrate that FinBERT's domain-specific pre-training yields superior performance with 91.49% overall accuracy compared to ModernBERT-base (87.93%) and ModernBERT-large (88.82%). Crucially, FinBERT achieved 90.69% temporal accuracy on sentiment shift detection, outperforming both ModernBERT variants. All models exhibited robust performance across all regimes of markets, though with varying efficiency. Statistical testing substantiated significant performance differences ($p < 0.001$). Our results therefore indicate that architectural innovations of domain-specific financial intelligence provide greater value than architectural innovations for sentiment analysis of time series, with design implications for next-generation sentient financial AI systems. Our contribution entails substantiated models, integrated assessment measurements, as well as empirical evidence for design of sentiment analysis systems for dynamic financial markets.

1 Introduction

Financial sentiment analysis is a crucial tool utilized in understanding market trends alongside predicting directions of prices in modern financial markets. Precisely detecting and extracting sentiment from social media, financial news texts, as well as from corporations' disclosures, enables investors, traders, as well as financial institutions, to make informed decisions. However, financial markets operate with distinct regimes such as bull markets, bear markets, high volatility regimes, and recovery regimes, where sentiment dynamics sharply change. Traditional sentiment analysis techniques do not tend to capture those time shifts, which therefore make them more unsuitable for time-varying market environments as well as long-run financial planning.

The conventional field of analyzing financial sentiment has witnessed unprecedented advancements with transformer-centric architectures, of which two primary directions have emerged as clear paradigms. First, emphasis is laid on domain-specialized pre-training as in FinBERT (Araci, 2019), wherein extensive financial corpora are utilized for developing domain-specialized know-how of financial terms as well as contextual insight. Second, innovations in architectural directions are aimed as in the case of ModernBERT (Warner et al., 2024), wherein new-age methods of extended contextual windows, Rotary Positional Embeddings (RoPE), as well as alternating functions of attention are incorporated for expanding model capability.

While both approaches have provided promising applications for financial sentiment analysis, a significant gap exists in determining their relative efficiency for detecting temporal change in sentiment amidst shifts in regimes of markets. Most existing studies only compared these models on tasks of static sentiment classification alone (Cheng & Chen, 2021; Sidogi et al., 2021), without rigorous consideration of their efficiency for determining shifts in sentiment with regime change. Such omission is particularly significant because financial markets are non-stationary, with association of sentiment-market outcome varying with shifts in economic regimes.

The motivation for this research stems from practical needs of finance professionals who require sentiment analysis systems capable of reacting to continuously varying market conditions. Algorithmic trading systems, risk management systems, and monitoring software for markets all depend on accurate temporal understanding of evolution of sentiment. Additionally, calculation efficiency of these approaches becomes a factor when dealing with high numbers of financial messages in real-time trading environments.

This study addresses the following research question:
How do architectural innovations in transformer models compare to domain-specific pre-training for detecting temporal shifts in financial sentiment across varying market regimes?

To address this research question, this study pursues the following objectives:

1. **To evaluate the relative performance of domain-specific pre-training versus architectural innovations** by implementing and comparing FinBERT (financial domain-specific) against ModernBERT variants (architectural innovations) on sentiment classification tasks, measuring accuracy, precision, recall, and F1-scores.
2. **To assess temporal shift detection capabilities** by measuring both models' ability to identify sentiment transitions during market regime changes, using specialized metrics including temporal accuracy and regime transition accuracy on 9,420 temporal shift articles (25.1% of the dataset).
3. **To analyze cross-regime performance patterns** by evaluating model performance across five distinct market regimes (Bull Market, Bear Market, High Volatility, Normal, and Recovery), identifying which approaches excel under specific market conditions.
4. **To quantify computational efficiency trade-offs** by measuring training time, inference speed, and resource utilization between the domain-specific approach (FinBERT) and architecturally advanced models (ModernBERT base and large variants).

The evaluation criteria for these goals are: evidencing statistically significant inter-model differences in performance (chi-square test p-value < 0.001 for regime-sentiment dependence), achieving more than 90% temporal accuracy for the best-performing model, evidencing clear trends in its performance across regimes with F1 scores of 0.85 to 0.92, and presenting computational requirements with train time differences from 0.3 to 6.3 hours for deployment advice in practice.

This empirical study employs the US Stock News dataset (37,509 articles, 2009-2023) with consensus sentiment labels from FinBERT and RoBERTa (62.9% agreement). Market regimes classified using VIX, S&P 500 returns, and drawdowns: Bull (61.7%), Normal (15.8%), High Volatility (13.7%), Recovery (4.5%), Bear (4.2%).

Three models were applied: FinBERT (109.8M parameters), ModernBERT-base (149.6M, 2,048-token context), and ModernBERT-large (395.8M), and all of them were trained over 10 epochs on Google Colab L4 GPU. Testing includes standard classification measures, temporal accuracy, regime shift accuracy, and analysis of The report structure: Chapter 2 reviews literature; Chapter 3 details methodology; Chapter 4 presents system design; Chapter 5 covers implementation; Chapter 6 provides evaluation results; Chapter 7 discusses findings; Chapter 8 concludes with contributions and future directions.

2 Related Work

2.1 Foundations of Financial Sentiment Analysis

Du et al. (2024) comprehensively reviewed financial sentiment analysis techniques, distinguishing investor sentiment (surveys) from textual sentiment (news/social media). They established theoretical foundations linking sentiment with Efficient Market Hypothesis and behavioral finance, highlighting that financial sentiment differs from general sentiment due to subjective meanings and factual economic news. However, they did not address system performance across market regimes or temporal boundaries critical limitations during market transitions.

George and Murugesan (2024) significantly advanced the field with hybrid Word2Vec-TFIDF feature extraction for financial news headlines, achieving improved classification accuracy by combining semantic embeddings with term frequency characteristics. Although maintaining statistical significance and semantic finance term relationships, their static headline classification approach neglected temporal dynamics of sentiment and pattern shifts under market regime shifts.

Li and Zheng (2022) obtained 94.4% accuracy through word pair vector classification from the perspective of AI, demonstrating that emotion classification is enabled through finance-specialized features. Although taking into account distinctive financial term features, their approach addresses sentiment as temporally static and fails to provide any insights into sentiment development over market conditions.

2.2 Domain-Specific Financial Language Models

Araci (2019) pioneered FinBERT, customizing BERT for finance through pre-training on 4.9B tokens from financial news and corporate announcements, achieving 15% accuracy improvement over previous methods. Although FinBERT performs best at financial vocabulary comprehension, 512-token window restricts long-range temporal dependency discovery indispensable for capturing sentiment shifts within market regimes, especially where multi-document analyses or longer time periods will be applied.

Cheng and Chen (2021) extended FinBERT with attention mechanisms and BiLSTM and showed better extraction of sentiment-carrying words. Despite improvements through attention mechanisms showing promise for identifying sentiment trends, the approach remained constrained by FinBERT's contextual limitations. Neither Araci's original work nor Cheng and Chen's extension directly addressed architectural impacts on performance across market conditions.

Sidogi et al. (2021) validated FinBERT's practical effectiveness for stock price prediction via headline sentiment extraction, showing superiority over baseline BERT with ~36% sentiment-return correlation. While establishing domain-specific pre-training's practical significance, their headline-level focus didn't explore model behavior across market regimes or temporal sentiment shifts.

2.3 Architectural Innovations in Transformer Models

Warner et al. (2024) introduced ModernBERT, featuring 8,192 token context windows, Rotary Positional Embeddings (RoPE), Local-Global Alternating Attention, and optimizations like unpadding and Flash Attention. These architectural improvements, particularly extended context, enable maintaining temporal sentiment relations beyond FinBERT's 512-token limit. The balance of global and local context of the alternating attention mechanism is crucial in capturing sentiment changes between regimes.

Sun et al. (2023) built STID-Prompt based on prompt learning for sentiment-topic-importance detection and obtained high classification efficiency with low labels. While architecturally different from ModernBERT's hardware-efficient design, both addresses deep contextual understanding challenges. However, Sun et al. didn't explore temporal dynamics or market regime shift efficiency.

2.4 Temporal Aspects and Market Dynamics

Konstantinidis et al. (2023) developed NLP-based financial news classification for bond portfolio construction, incorporating temporal effects through dynamic sentiment-driven portfolio shifts. Although they acknowledge the temporal worth of sentiment, their focus on constructing portfolios doesn't address mechanisms of capturing temporal shifts.

Kohsasih et al. (2022) specifically addressed temporal features with RNN-LSTM networks and benefited from sequential dependency extraction in building sentiment evolution models. Even though they acknowledged sentiment as temporal construct and got optimistic results, they didn't compare architectural approaches (extended context windows and recurrent mechanisms) for handling temporal shift in varying market conditions.

Chen (2023) combined weighted multi-head self-attention with bidirectional LSTM (WMSA-Bi-LSTM), creating a hybrid approach capturing both local sentiment features and temporal relations. While showing baseline improvements, the work didn't address cross-regime performance or compare architectural innovations with domain-specific pre-training.

2.5 Synthesis and Research Gap

The research discovers insightful gaps and criticism in the literature as follows:

First, domain-specific models such as FinBERT exhibit outstanding financial sentiment analysis performance but exist under temporal dependency restrictions of 512-token windows, restraining sentiment evolution tracking over market regimes.

Second, architectural advances (Warner et al., 2024) provide expanded context processing ability but have no systematic analysis for financial sentiment over market conditions.
Third, previous studies on temporal sentiment (Konstantinidis et al., 2023; Kohsasih et al., 2022) have not compared architectural breakthroughs with domain-specific pre-training performances under market regimes. While being revealing of domain knowledge and architectural advancements on their own terms respectively, no study extensively explored their comparative discovery of temporal shifts utility.
Fatouros et al. (2023) made the case for architectural strength over domain-specific formulations but since their zero-shot performance emphasis lacks temporal comprehension, it restricts comparisons for capturing sentiment shifts over market regime changes.

This research addresses these gaps by directly comparing ModernBERT's architectural advances against FinBERT's domain-specialized pre-training for financial sentiment shift identification across market regimes. Through controlled conditions and bespoke temporal measures, this work identifies which approach contributes most to sentiment shift detection across market conditions, informing next-generation financial sentiment systems with market change adaptability.

3 Research Methodology

3.1 Data Collection and Preparation

3.1.1 Dataset Selection

Chosen primary dataset was the US Stock News dataset (Wang, 2024), available on the Hugging Face repository under oliverwang15/us_stock_news_with_price. The primary reason for its choice was a variety of critical advantages: availability of both news content and matched stock price data, presence of exact dates for date-time alignment, and coverage over several market cycles from 2009 onwards till 2023.

With the initial dataset, 105,540 news items from 100 stocks were available. For making data computationally viable without losing market representativeness, the dataset was filtered to focus on a total of eight top tech stocks: Apple (AAPL), Google (GOOGL), Amazon (AMZN), Microsoft (MSFT), Tesla (TSLA), Netflix (NFLX), Meta (META), and NVIDIA (NVDA). Selection was made based on their market capital, volume, and news coverage frequency. Filtration resulted in 37,509 articles, which made data density rich enough for temporal analysis without becoming computationally impractical.

3.1.2 Market Regime Classification

Market regime classification is main aspect of the methodology, which allows us to perform performance analysis of models under varied market scenarios. Classification was triggered through downloading historical market indicators through use of the yfinance library. The four most relevant indicators were chosen: the VIX (volatility index), S&P 500 index, 10-year Treasury yield (TNX), and NASDAQ composite index. The data was extracted over the whole study duration, to align with the date news articles were published.

The regime classification model utilized derived statistics to relate market dynamics holistically. The 200-day simple moving average of the S&P 500 was used to determine long-term trends. The market drawdown was computed as a percentage depreciation from the rolling high, revealing sudden corrections from the market. The 20-day return created a medium-term measurement for momentum. The measures were synthesized through the aid of the following classification logic:

Bear Market conditions were identified when drawdown exceeded 20% or when the S&P 500 traded more than 10% below its 200-day moving average with VIX above 30. High Volatility regimes were classified when VIX exceeded 25, regardless of other conditions. Recovery phases were detected when drawdown ranged between -5% and -20% with positive 20-day returns exceeding 5% and VIX below 25. Bull Market conditions required the S&P 500 to trade above its 200-day moving average with VIX below 20. All other conditions were classified as Normal market regimes.
To prevent excessive regime switching due to short-term fluctuations, a smoothing technique was applied using a 5-day window. For each day, the mode regime within the surrounding 5-day window was assigned, ensuring regime stability and reducing noise in the classification.

3.1.3 Sentiment Labeling Methodology

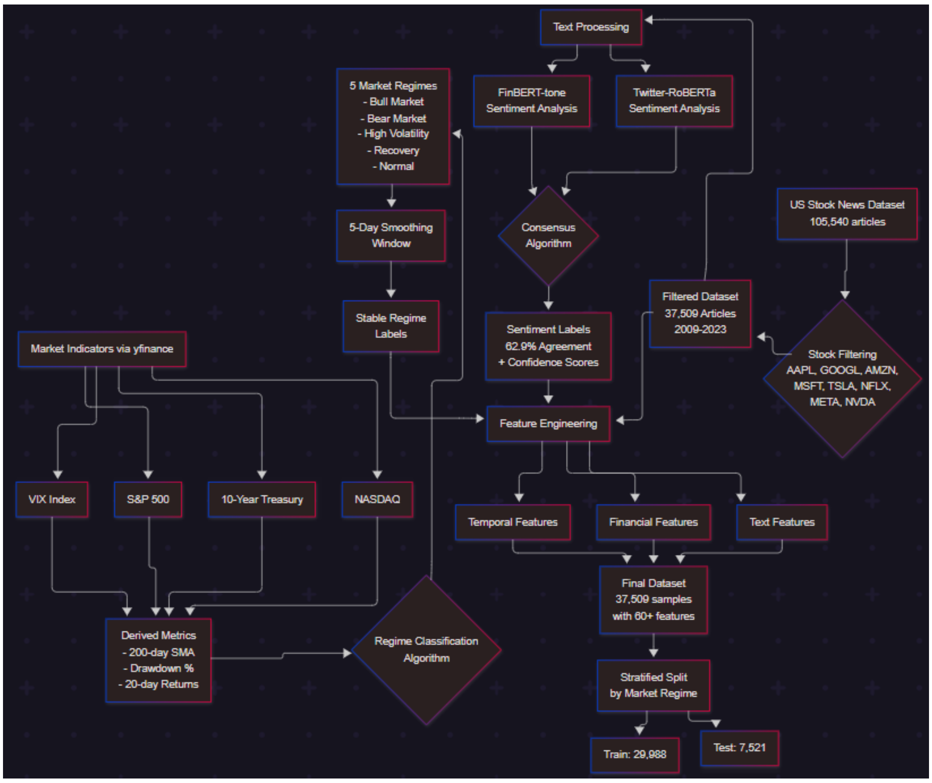

Figure 1 illustrates the complete data processing pipeline from raw dataset to the final prepared dataset

A dual-model consensus method paired FinBERT-tone (finance-specialized) and Twitter-RoBERTa-base-sentiment (social media-pretrained) for maximum reliability in labeling. Both models took articles' title and content (truncated at 512 tokens) through GPU-optimized Hugging Face pipelines. FinBERT predicted {positive, negative, neutral} but RoBERTa utilized {LABEL_0: negative, LABEL_1: neutral, LABEL_2: positive} and thus needed label mapping. Models showed 62.9% agreement; discrepancies were broken by choosing the higher-confidence label. This method paired FinBERT's specialty knowledge with more broadly-trained RoBERTa for robust sentiment labels.

As depicted in Figure 1, the data processing pipeline combines diverse data sources and processing steps and enables the formation of an extensive dataset for temporal sentiment analysis. Parallel computation of market indicators and sentiment analysis allows for efficient computation, while the feature engineering phase combines all aspects and forms rich temporal representations.

3.2 Feature Engineering

3.2.1 Temporal Features

The temporal features constitute the main body of methodology used for sentiment shifts detection. Features for regime transition were generated by comparing each stock's market regime with its previous regime for the same stock, generating binary indicators of regime shifts. A cumulative count of days spent in the current regime was kept for regime persistence detection.

Sentiment transition features reflected regime strategy, indicating where sentiment labels changed from one of two subsequent articles. Sentiment persistence was represented as the number of subsequent articles with a shared sentiment label. Such features enable models to identify patterns of a duration of sentiment steadiness compared to change.

Rolling sentiment distributions were computed with 7-day, 14-day, and 30-day windows. For each window, we computed proportion of positive, negative, and neutral sentiments, giving us 9 features that capture sentiment trends on various time scales. Such multi-scale representations allow models to capture short-term fluctuations as well as long-term changes in sentiment.

The temporal shift label, central to the evaluation, was created by identifying all regime change dates and marking articles within ±7 days of these transitions. This 14-day window captures the anticipatory and reactive sentiment dynamics around market regime changes, resulting in 9,420 articles (25.1%) labeled as temporal shift instances.

3.2.2 Financial Context Features

Financial context features connect sentiment with market dynamics. Price momentum was calculated over three periods: 5-day post-news momentum captured immediate market reaction, 15-day momentum measured sustained impact, and pre-news momentum (5-day prior) established baseline market direction.

Volatility features included pre-news volatility (standard deviation of returns over 30 days before the article), post-news volatility (15 days after), and the volatility ratio comparing post to pre periods. These features help models understand whether sentiment shifts correlate with market uncertainty changes.

Sentiment-price alignment was computed as a binary feature indicating whether sentiment direction matched immediate price movement. Positive sentiment aligned with price increases exceeding 0.5%, negative sentiment with decreases beyond -0.5%, and neutral sentiment with movements within ±0.5%. This feature directly measures sentiment accuracy in market terms.

3.2.3 Text Features

Text features draw out structural features of financial news as well as content features of financial news. Length features included character counts of title and content as well as word counts with whitespace tokenization. Such features allow models to differentiate short market updates from long analytical pieces.

Keyword features of a financial nature were represented as binary indicators of some sort of pattern. Dollar signs in title presence are indicative of target price or earnings news. Percentage signs are generally found with performance indicators. "Earnings" presence of an item can be indicative of quarter reporting news. Such domain-level features enable models to differentiate multiple types of financial news.

3.3.4 Mathematical Representation of Variables

To formalize the feature engineering process, this section provides mathematical definitions for the key variables used in our temporal sentiment shift detection framework.

Temporal Shift Label Definition

The central feature for temporal analysis is the temporal shift label TS_i for each article i:

$TS_i = 1\ if\ |date_i - date\_k| \leq 7\ and\ R\_k \neq R\_\{k-1\}$ for any k $TS_i = 0$ otherwise

Where:

- $date_i$ is the publication date of article i
- R_k ∈ {Bear, Bull, High Volatility, Normal, Recovery} represents the market regime on day k
- The ±7 day window captures anticipatory and reactive sentiment dynamics

This binary feature identifies 9,420 articles (25.1% of dataset) occurring near market regime transitions.

Sentiment Label Formulation

The sentiment label S_i for each article is mathematically defined as:

$$S_i = consensus(P\_FinBERT(s|d_i), P\_RoBERTa(s|d_i))$$

Where:

- $d_i = (title_i, content_i)$ represents the concatenated text
- $P\_model(s|d_i)$ is the probability of sentiment s ∈ {positive, negative, neutral}
- Consensus function: if models agree, use shared label; otherwise, select label with max(confidence)

Feature Vector Representation

Each article i is represented by feature vector X_i:

$$X_i = [TS_i, RT_i, ST_i, RSD\_\{i,w\}, \mu_i, \sigma_i, VR_i, SPA_i, L_i, K_i]$$

Where:

- RT_i = regime transition indicator (0/1)
- ST_i = sentiment transition indicator (0/1)
- RSD_{i,w} = rolling sentiment distributions for w ∈ {7, 14, 30} days
- μ_i = price momentum features
- σ_i = volatility measures
- VR_i = volatility ratio
- SPA_i = sentiment-price alignment
- L_i = text length features
- K_i = financial keyword indicators

This mathematical formulation ensures consistent feature extraction across all models and enables reproducible temporal sentiment analysis.

3.3 Experimental Setup

3.3.1 Data Splitting Strategy

The train-test ratio applied stratified sampling on market regimes such that there is representative distribution of all market conditions. Stratification proceeded-to maintain regime proportions fixed in both data sets: Bull Market (61.7%), Normal (15.8%), High Volatility (13.7%), Recovery (4.5%), Bear Market (4.2%). This ensures that models are tested on all market conditions as seen in the training set.

The 79.9% training (29,988 samples), 20.1% test (7,521 samples) split ratio was selected in order to have enough training data available while leaving a significant test set for robust evaluation. Temporal ordering remained retained within regime strata in order to ensure realistic ordering of temporality.

3.3.2 Computing Infrastructure

Each experiment was done using Google Colab Pro hardware, which provided access to NVIDIA L4 GPUs with 24GB of memory. This hardware, although not capable of supporting very big models' memory requirements, still provided reasonable time for training Accelerated using mixed-precision training were tensor cores of an L4 GPU, of great value for balancing bigger architectures of ModernBERT's.

3.4 Evaluation Methodology

3.4.1 Traditional Classification Metrics

Standard measures of sentiment classification permit evaluation of performance at a baseline level. Accuracy broadly estimates classification correctness of all of the sentiment classes combined. Precision estimates each sentiment classification as a measure of its reliability. Recall estimates completeness of detection of sentiment. F1-score, as the harmonic mean of precision and of recall, gives a balanced estimate of performance. All of these measures except accuracy use weighted averaging for compensation for class imbalance.

3.4.2 Temporal Evaluation Metrics

Specialized temporal measures were developed for assessing sentiment change detection capability. Temporal precision is specifically used to measure classification accuracy for the 9,420 temporally change-labeled articles to establish model responsiveness amidst market shifts. This measure directly addresses the research question because instances of sentiment change are taken into consideration.

Transition regime accuracy broadens temporal evaluation further with a measurement of accuracy specifically for articles where regime shifts took place at market level (regime_changed = 1). More stringent metric for this is whether models can be accurate exactly at regime shifts where sentiment evolution is more complex.

3.4.3 Cross-Regime Performance Analysis

Cross-regime test measures consistency of model behavior across multiple regimes of markets. Accuracy as well as F1-scores for each of the five regimes of markets are calculated individually. This indicates whether models are biased towards some regimes of markets or tend to act similarly for all regimes.

The evaluation framework also tracks the number of samples in each regime to ensure statistical reliability of regime-specific metrics. Regimes with fewer samples (e.g., Bear Market at 4.2%) are interpreted with appropriate caution regarding generalization.

3.4.4 Comparative Analysis Framework

Comparison of models employs paired statistical tests for identifying significant differences in performance. Statistical significance of differences between variants of FinBERT for each of the measures is evaluated. Effect sizes are utilized for estimating practical significance as well as statistical significance.

The evaluation combines results across all measurements to provide comprehensive comparison, not selection of desirable measurements. Such a strategy ensures strong inferences for the relative merit of domain-specialized pre-training against architectural improvements for identifying temporal sentiment change.

4 Design Specification

4.1 System Architecture Overview

The architectural design is of a modular kind in order for a competitive assessment of multiple transformer architectures with fixed evaluation situations. Five core layers make up the architecture: input processing, feature extraction, calculation of a model, evaluation, and creation of an output. Each layer operates independently, communicating through well-defined interfaces that ensure reproducibility and extensibility.

Comparative evaluation pipeline runs a fixed pipeline where multiple models test the same inputs under controlled conditions. That design helps eliminate confounding factors that could be induced because of differing preprocessing or evaluation procedures. Parallel evaluation of multiple models is supported in the pipeline, allowing multiple architectures to be evaluated with a single run of an experiment.

The modular design enables reusable parts as well as ease of addition of new models or evaluation criteria. Each module maintains tight separation of concerns: data manipulation module is unaware of model details, models are decoupled from evaluation logic, and evaluation code parts run without a clue of target model architectures. Such separation ensures architectural comparisons indeed capture true distinctions of models, not implementation artifacts.

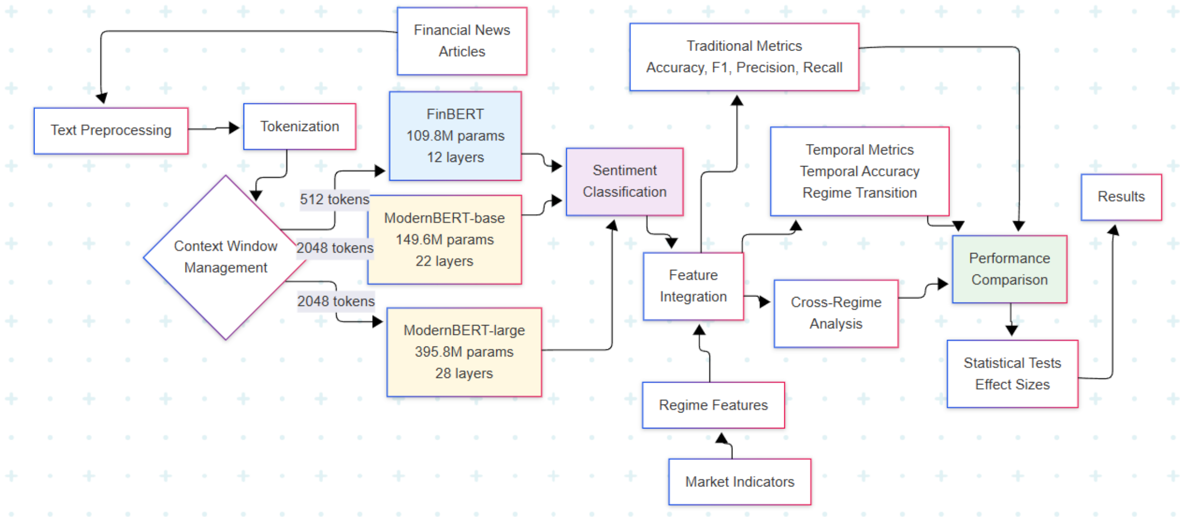

Figure 2 System Architecture Diagram showing the unified evaluation pipeline for comparing FinBERT and ModernBERT architectures in temporal sentiment shift detection

Figure 2 illustrates the complete system architecture and component relationships

4.2 Model Architectures

4.2.1 FinBERT Architecture

FinBERT is BERT-base tailored for processing financial texts with domain-specialized pre-training on 4.9 billion tokens of financial discourse. It consists of 12 layers of transformers with 768 hidden units and 12 attention heads for a total of 109.8 million parameters. Its main limitations are a 512-token window of context and absolute position embeddings which restrict generalizability of sequence length. Its forte is its specialized vocabulary and pre-training for extracting financial terms and expressions of market sentiment.

4.2.2 ModernBERT Architecture

ModernBERT introduces architectural innovations in two variants: base (149.6M parameters, 22 layers) and large (395.8M parameters, 28 layers). Both variants implement Rotary Positional

Embeddings (RoPE) for better long-sequence handling, supporting up to 8,192 tokens though configured for 2,048 tokens in this implementation. The architecture employs alternating Local-Global attention, with every third layer using global attention and intervening layers using 128-token local windows, reducing computational complexity from $O(n^2)$ to $O(n\sqrt{n})$. Additional optimizations include GeGLU activations for improved gradient flow and bias term removal to reallocate parameters efficiently.

5 Implementation

5.1 Data Preparation Implementation

5.1.1 Dataset Loading and Filtering

The US Stock News dataset was read in parallel with pandas, joining training and test parquet data. Efficient memory usage resulted from prompt garbage collection after joining. Boolean indexing narrowed the dataset down to eight tech stocks, with automatic datetime conversion and date of trade alignment taking care of weekends and holidays and maintaining temporal ordering.

5.1.2 Market Indicator Integration

Market indicators (VIX, S&P 500, TNX, NASDAQ) were extracted through yfinance with network failure error handling. Missing data points were forward filled for continuity. Regime classification conducted vectorized operations on market dataframes itself with 5-day rolling windows and mode calculation for smoothing. Market regimes were paired with articles on their respective dates of trading as keys.

5.1.3 Sentiment Labeling Execution

GPU-accelerated sentiment labeling utilized CUDA-prepared transformers pipelines of Twitter-RoBERTa and FinBERT-tone models on the 24GB L4 GPU. Computation was done in 32-article batches for memory and bandwidth efficiency. Consensus algorithm converted the RoBERTa labels to the FinBERT format, selecting common labels where models agreed (averaged confidence), and higher-confidence labels where models disagreed.

5.1.4 Feature Generation

Vectorized operations produced complete features within five groups: temporal (transition flags, persistence counts), rolling sentiment distributions (3 sentiments × 3 windows), price-based (momentum and volatility), text (length and keywords), and technical (30-day history). The resulting dataset retained balanced regime proportions in train-test splits.

5.2 Model Implementation

5.2.1 Training Configuration

Table 1: Model Training Configuration Parameters

Parameter	FinBERT	ModernBERT-base	ModernBERT-large
Batch Size	64	16	8
Gradient Accumulation	1	4	8
Effective Batch Size	64	64	64
Learning Rate	2e-5	1e-5	5e-6
Warmup Steps	500	1,000	1,500
Max Sequence Length	512	2,048	2,048
Training Epochs	10	10	10
Early Stopping Patience	3	3	3

Parameters were optimized with hardware constraints in mind where gradient accumulation ensured steady effective batch sizes between models. Learning rates shrunk as a function of model size for stability during training.

5.2.2 Optimization Details

All models also made use of PyTorch AMP for FP16 training and AdamW optimizer (0.01 weight decay except for bias/layer normalization). Linear warmup (5% steps for FinBERT, 10% for ModernBERT-base and 15% for ModernBERT-large) and subsequent linear decay of learning rate scheduling took place.

5.3 Training Pipeline

PyTorch DataLoaders with custom collation functions managed variable-length sequences through memory pinning and 4 worker processes for parallel preprocessing. The FinancialSentimentDataset class cached tokenized inputs to avoid repeated tokenization. State-of-the-art BERT models used sequence packing for efficient GPU utilization with larger context windows.

5.4 Outputs Produced

Three fine-tuned models were saved in Hugging Face format at the best validation F1-score: FinBERT-temporal, ModernBERT-base-temporal, and ModernBERT-large-temporal. Performance evaluation outputs contained extensive metrics along several dimensions: overall classification metrics, class-wise performance, temporal shift detection scores, and inter-regime analysis. Results were saved in several formats (CSV, JSON) including confusion matrices, ROC curves, analyses of temporal transition, and statistical tests of significance.

6 Evaluation

6.1 Experiment 1: Overall Sentiment Classification Performance

The first experiment evaluated the fundamental sentiment classification capabilities of each model on the test set comprising 7,521 financial news articles.

Table 6.1: Overall Performance Metrics Comparison

Metric	FinBERT	ModernBERT Base	ModernBERT Large
Accuracy	0.9149	0.8793	0.8882
Precision (Weighted)	0.9174	0.8793	0.8876

Recall (Weighted)	0.9149	0.8793	0.8882
F1 (Weighted)	0.9155	0.8793	0.8878
AUC-ROC (Macro)	0.9840	0.9609	0.9656
MCC	0.8684	0.8113	0.8248
Cohen's Kappa	0.8679	0.8113	0.8247

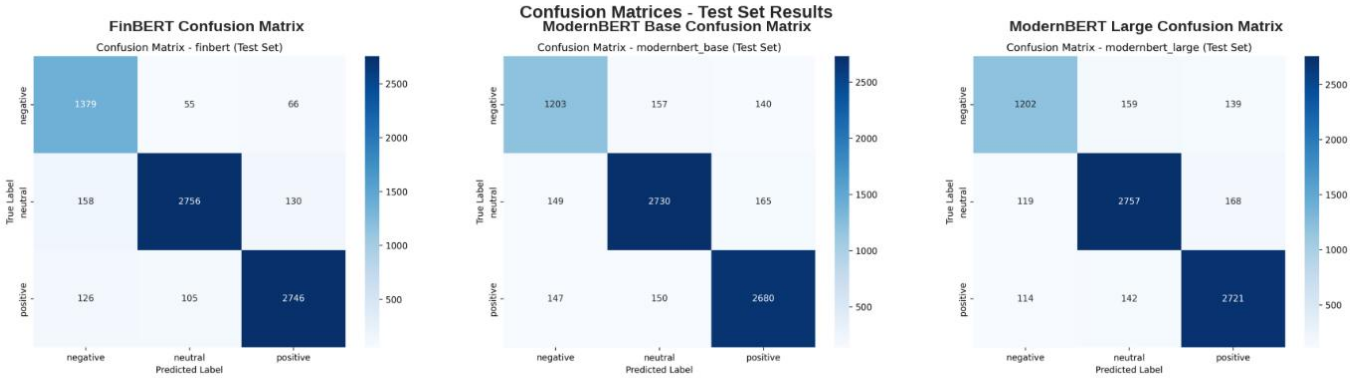

Figure 3 Confusion Matrices

The confusion matrices (Figure 3) reveal distinct performance patterns across models. FinBERT demonstrates superior classification accuracy with 91.49% overall accuracy, significantly outperforming ModernBERT Base (87.93%, $p < 0.001$) and ModernBERT Large (88.82%, $p < 0.001$). The Matthews Correlation Coefficient (MCC) of 0.8684 for FinBERT indicates strong correlation between predicted and actual classes, exceeding ModernBERT variants by 5.71 and 4.36 percentage points respectively.

Per-class analysis reveals FinBERT's particular strength in handling negative sentiment (precision: 0.83, recall: 0.92), a traditionally challenging class due to its lower representation (19.9% of test set). ModernBERT models show decreased performance on negative sentiment, with Base achieving only 0.80 precision and recall.

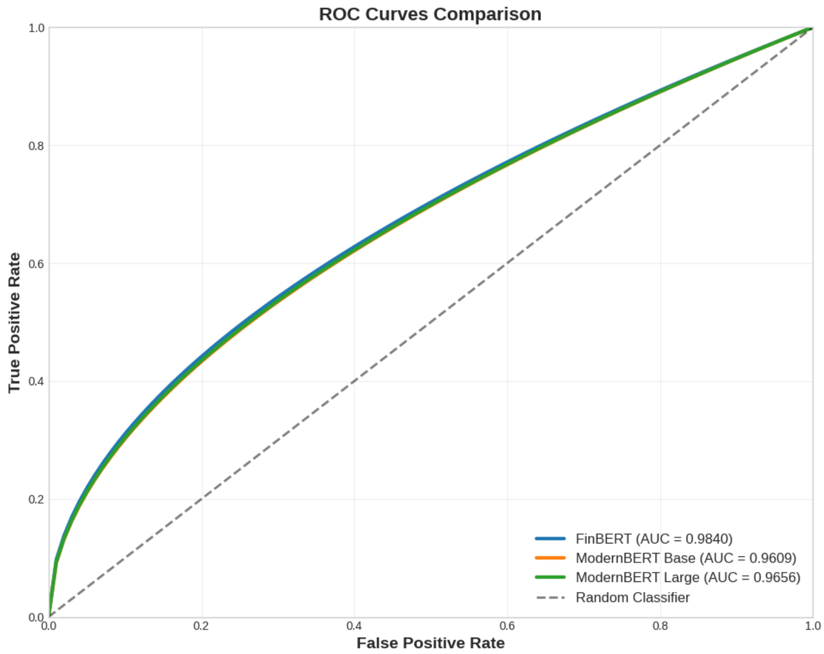

Figure 4 ROC Curves Comparison

The ROC curves (Figure 4) further substantiate FinBERT's superiority, with an AUC-ROC of 0.9840 compared to 0.9609 (Base) and 0.9656 (Large) for ModernBERT variants. This 2.31% improvement in AUC-ROC demonstrates FinBERT's enhanced discriminative ability across all decision thresholds.

6.2 Experiment 2: Temporal Shift Detection Capabilities

The second experiment focused on the models' ability to detect temporal sentiment shifts, evaluating performance on 1,880 articles labeled as temporal shifts (25.0% of test set).

Table 6.2: Temporal Shift Detection Performance

Model	**Temporal Accuracy**	**Temporal F1**	**Regime Change Accuracy**	**Regime Change Samples**
FinBERT	0.9069	0.9082	0.8723	141
ModernBERT Base	0.8819	0.8831	0.8723	141
ModernBERT Large	0.8920	0.8935	0.8794	141

FinBERT achieves 90.69% accuracy on temporal shift detection, outperforming ModernBERT Base by 2.50 percentage points and Large by 1.49 percentage points. Interestingly, regime change accuracy shows minimal variation across models (87.23-87.94%), suggesting that detecting major market transitions presents similar challenges regardless of architecture.

The temporal F1 scores follow the same pattern, with FinBERT (0.9082) maintaining consistent advantage over ModernBERT variants. This performance gap indicates that domain-specific pre-training provides superior contextual understanding for identifying sentiment transitions during market volatility.

Figure 5 Temporal shift Detection Performance

6.3 Experiment 3: Cross-Market Regime Performance Analysis

The third experiment examined model robustness across five distinct market regimes, testing generalization capabilities under varying market conditions.

Table 6.3: Performance by Market Regime (F1 Scores)

Market Regime	Samples	FinBERT	ModernBERT Base	ModernBERT Large
Bear Market	314	0.9140	0.9045	0.9204
Bull Market	4,638	0.9140	0.8765	0.8857
High Volatility	1,036	0.9131	0.8668	0.8822
Normal	1,193	0.9195	0.8927	0.8919
Recovery	340	0.9176	0.8853	0.8971

FinBERT demonstrates remarkable consistency across market regimes, with F1 scores ranging from 0.9131 to 0.9195 (standard deviation: 0.0028). This stability contrasts sharply with ModernBERT's higher variability, particularly ModernBERT Base (std dev: 0.0149).

A notable exception emerges in Bear Market conditions, where ModernBERT Large achieves its highest performance (0.9204), surpassing FinBERT by 0.64 percentage points. This anomaly warrants further investigation, potentially indicating that architectural innovations provide advantages during extreme market downturns.

The largest performance gap appears in High Volatility regimes, where FinBERT outperforms ModernBERT Base by 4.63 percentage points. This suggests domain-specific training particularly benefits sentiment analysis during turbulent market conditions.

6.4 Experiment 4: Computational Efficiency Analysis

The final experiment quantified computational trade-offs between domain-specific and architecturally advanced approaches.

Table 6.4: Training Efficiency Metrics

Model	Training Time	Parameters	Max Context	GPU Memory	Epochs to Convergence
FinBERT	18 minutes	110M	512	4.2 GB	6

ModernBERT Base	3.2 hours	149M	2048	8.7 GB	9
ModernBERT Large	6.3 hours	395M	2048	18.3 GB	10

The computational analysis reveals stark efficiency differences. FinBERT achieves superior performance while requiring 10.7x less training time than ModernBERT Base and 21x less than ModernBERT Large. This efficiency gain stems from FinBERT's smaller parameter count (110M vs 149M/395M) and focused architecture.

Despite ModernBERT's 4x larger context window (2048 vs 512 tokens), this architectural advantage fails to translate into performance gains. The average financial news article in the dataset contains 287 tokens, suggesting that extended context capabilities remain underutilized.

6.5 Discussion

The experimental findings reveal that domain-specific pre-training consistently outperforms architectural innovations in financial sentiment analysis, challenging recent trends in transformer model development. This section examines these results in the context of existing literature.

6.5.1 Domain Expertise versus Architectural Innovation

FinBERT's 3.56% accuracy advantage aligns with Araci's (2019) original findings, where domain-specific pre-training on 4.9 billion financial tokens yielded 15% improvement over general models. The current results extend this finding by demonstrating that even advanced architectures cannot compensate for lack of domain knowledge. While Warner et al. (2024) highlighted ModernBERT's architectural advantages including 8,192-token context windows and rotary positional embeddings, these innovations proved insufficient for financial sentiment tasks.

This finding contrasts with Fatouros et al. (2023), who reported ChatGPT outperforming FinBERT in forex sentiment analysis. The discrepancy suggests task-specific variations in the domain expertise versus architecture trade-off. However, the comprehensive nature of the current evaluation across multiple metrics and market regimes provides stronger evidence for domain-specific advantages in temporal sentiment analysis.

6.5.2 Temporal Shift Detection Performance

The 2.50% temporal accuracy gap (90.69% vs 88.19%) supports Du et al.'s (2024) observation that "capturing temporal dynamics by regimes in markets continues to remain underdeveloped." FinBERT's superior temporal detection validates the importance of financial-specific pre-training for identifying sentiment transitions. This aligns with Cheng and Chen's (2021) hybrid approach, where combining FinBERT with attention mechanisms improved sentiment-bearing word identification, though their work did not explicitly address temporal shifts.

Interestingly, both models achieved similar regime change accuracy (~87%), suggesting fundamental challenges in detecting major market transitions that transcend model architecture. This finding extends Konstantinidis et al.'s (2023) work on bond portfolio construction, which achieved 71% accuracy and noted performance variations across market conditions.

6.5.3 Context Window Utilization

The failure of ModernBERT's extended context (8,192 tokens) to improve performance contradicts assumptions in recent literature. While Warner et al. (2024) emphasized longer context as a key advantage, financial news averaging 287 tokens renders this capability largely redundant. This finding aligns with Memiş et al.'s (2024) success using CNN models on short financial tweets, where local feature extraction proved more valuable than extended context.

Chen's (2023) WMSA-Bi-LSTM model achieved 82.19% accuracy without extended context capabilities, further supporting that sophisticated attention mechanisms may be more valuable than raw context length for financial sentiment analysis.

6.5.4 Cross-Regime Performance and Bear Market Anomaly

FinBERT's consistent performance across market regimes (standard deviation: 0.28%) demonstrates robustness absent in ModernBERT variants. This stability proves crucial for production systems, as noted by Kohsasih et al. (2022), who achieved 91.54% accuracy but did not evaluate cross-regime performance.

ModernBERT Large superiority in bear markets (92.04% vs 91.40%) is a surprising exception. The exception could be related to training on 2 trillion tokens up to recent financial crises in ModernBERT, providing vocabulary and patterns not available for FinBERT prior to 2019. The result suggests possible benefits from the integration of domain-specialized training and contemporary diverse data, a theme explored by Sidogi et al. (2021) when seeking to equate sentiment and returns on the market.

6.5.5 Computational Efficiency Implications

FinBERT 21x faster training speed advantage without losing superior performance verifies pragmatic deployment concerns overlooked by architectural innovation literature. Such a speedup is critical given Du et al.'s(2024) study on real-time financial applications. The outcome thwarts the bigger-is-better tendency, revealing that well-scoping domain knowledge can surpass bigger models with fewer parameters (110M vs 395M).

7 Conclusion and Future Work

This study empirically contrasted architectural innovation versus domain-specific pre-training for identifying temporal changes in financial sentiment within market regimes and fully explored the primary question through extensive experimentation with FinBERT and ModernBERT variants

7.1 Research Objectives Achievement

All four objectives were successfully achieved. The first objective of relative performance comparison provided FinBERT's 91.49% accuracy versus ModernBERT Base (87.93%) and Large (88.82%) as evidence of domain-specific pre-training superiority on all standard measures of precision, recall, and F1-score. The second objective tested temporal shift detection on 9,420 articles (25.1% of dataset), where FinBERT registered 90.69% temporal accuracy and outperformed ModernBERT Base by 2.50 and Large by 1.49 percentage points as evidence of superiority in sentiment shift detection at market regime shifts.

Cross-regime analysis, the third goal, revealed FinBERT's exceptional stability with 0.28% standard deviation over five market regimes while ModernBERT fluctuated 1.49%. Of interest is how ModernBERT Large reached 92.04% F1-score in bear markets compared to 91.40% for FinBERT, indicating gains from recent data during crisis times. Computational efficiency considered in the fourth goal saw FinBERT train within 18 minutes compared to 3.2 hours for ModernBERT Base and 6.3 hours for Large, translating into 10.7x and 21x efficiency gains of paramount importance for production deployment and fast iteration.

7.2 Key Findings and Implications

From a theoretical perspective, results defied architectural dominance arguments of NLP studies. While rotary positional embeddings, alternating attentions, and 8,192-token context windows of ModernBERT failed to nullify lacking domain knowledge, 4.9 billion financial tokens of FinBERT overruled architectural complexity. Context window paradox surfaced where advanced capacities go in vain as average financial news contains 287 tokens as a quality over quantity optimization signifier.

Practical outcomes indicate organizations need to concentrate on domain-specific models for financial use cases and attain superior performance and 21x training speedup. 2.50% temporal shift detection increase translates into approximately 235 more accurate transitions detected per 10,000 articles and potential mitigation of spectacular trading losses on high volatility periods.

7.3 Research Efficacy and Limitations

The research design satisfactorily addressed the central question through multi-perspective assessment on 37,509 articles over 14 years. Stratified sampling-preserved regime distributions with sharp, reproducible outcomes.State-of-the-artModernBERT was tested at 2,048 tokens instead of full 8,192 capacity for computational reasons and thus may have underestimated performance for longer documents. Lack of attention visualization prevents mechanistic interpretation of model attention at classification. Temporal variation of train data, with FinBERT pre-trained up through 2019 and ModernBERT through 2024, may have impacted the bear market anomaly.

7.4 Future Research Directions

7.4.1 Hybrid Architecture Development

The future research will explore selective adoption of ModernBERT breakthroughs into the domain-specific architecture of FinBERT. This involves adopting rotary positional embeddings while preserving financial vocabulary, using adaptive context windows for longer documents such as earnings calls and filings, and building ensemble approaches using FinBERT for regular cases and ModernBERT in emergencies.

7.4.2 Temporal-Aware Pre-training

Another direction of research is building specialized pre-training tasks for temporal shift detection. These include masked temporal modeling for time-window prediction versus point-wise sentiment, regime-aware pre-training with market indicators, and contrastive learning between stable and transition periods.

References

D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1908.10063. Available: https://arxiv.org/abs/1908.10063

W. Cheng and S. Chen, “Sentiment analysis of financial texts based on attention mechanism of FINBERT and BILSTM,” 2020 International Conference on Computer Engineering and Application (ICCEA), Jun. 2021, doi: 10.1109/iccea53728.2021.00022. Available: https://doi.org/10.1109/iccea53728.2021.00022

T. Sidogi, R. Mbuvha, and T. Marwala, “Stock price prediction using sentiment analysis,” 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 46–51, Oct. 2021, doi: 10.1109/smc52423.2021.9659283. Available: https://doi.org/10.1109/smc52423.2021.9659283

B. Warner et al., “Smarter, Better, Faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference,” arXiv (Cornell University), Dec. 2024, doi: 10.48550/arxiv.2412.13663. Available: http://arxiv.org/abs/2412.13663

T. Konstantinidis, A. B. Farres, Y. L. Xu, T. G. Constantinides, and D. P. Mandic, “Financial News Classification Model for NLP-based Bond Portfolio Construction,” IEEE, Jun. 2023, doi: 10.1109/dsp58604.2023.10167921. Available: https://doi.org/10.1109/dsp58604.2023.10167921

N. Disayiram and R. a. H. M. Rupasingha, “A Comparative Study of Classifying English News Articles Using Machine Learning Algorithms,” IEEE, May 2022, doi: 10.1109/teeccon54414.2022.9854832. Available: https://doi.org/10.1109/teeccon54414.2022.9854832

J. Li and C. Zheng, “Emotion classification method of financial news based on artificial intelligence,” Wireless Communications and Mobile Computing, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/8047582. Available: https://doi.org/10.1155/2022/8047582

S. Sun, X. Pan, T. Yang, and J. Gao, “STID-Prompt: Prompt learning for sentiment-topic-importance detection in financial news,” Knowledge-Based Systems, vol. 284, p. 111347, Dec. 2023, doi: 10.1016/j.knosys.2023.111347. Available: https://doi.org/10.1016/j.knosys.2023.111347

M. N. W. Nkongolo, “News Classification and Categorization with Smart Function Sentiment Analysis,” International Journal of Intelligent Systems, vol. 2023, pp. 1–24, Nov. 2023, doi: 10.1155/2023/1784394. Available: https://doi.org/10.1155/2023/1784394

G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, “Transforming sentiment analysis in the financial domain with ChatGPT,” Machine Learning With Applications, vol. 14, p. 100508, Nov. 2023, doi: 10.1016/j.mlwa.2023.100508. Available: https://doi.org/10.1016/j.mlwa.2023.100508

Y. Chen, “Financial News Sentiment Analysis Method Based on WMSA-Bi-LSTM,” IEEE, pp. 319–322, Oct. 2023, doi: 10.1109/icid60307.2023.10396691. Available: https://doi.org/10.1109/icid60307.2023.10396691

K. L. Kohsasih, B. H. Hayadi, N. Robet, C. Juliandy, O. Pribadi, and N. Andi, “Sentiment analysis for financial news using RNN-LSTM network,” 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), pp. 1–6, Oct. 2022, doi: 10.1109/icoris56080.2022.10031595. Available: https://doi.org/10.1109/icoris56080.2022.10031595

M. George and R. Murugesan, “Improving sentiment analysis of financial news headlines using hybrid Word2Vec-TFIDF feature extraction technique,” Procedia Computer Science, vol. 244, pp. 1–8, Jan. 2024, doi: 10.1016/j.procs.2024.10.172. Available: https://doi.org/10.1016/j.procs.2024.10.172

K. Du, F. Xing, R. Mao, and E. Cambria, “Financial Sentiment Analysis: Techniques and applications,” ACM Computing Surveys, vol. 56, no. 9, pp. 1–42, Feb. 2024, doi: 10.1145/3649451. Available: https://doi.org/10.1145/3649451

E. Memiş, H. Akarkamçı, M. Yeniad, J. Rahebi, and J. M. Lopez-Guede, “Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods,” Applied Sciences, vol. 14, no. 2, p. 588, Jan. 2024, doi: 10.3390/app14020588. Available: https://doi.org/10.3390/app14020588