Report

Abstract

This report provides a comprehensive analysis of various machine learning architectures for predicting stock directional accuracy. Two main categories of models are evaluated: those focusing on time-series price data (1D Convolutional Neural Networks, Long Short-Term Memory networks, and Temporal Convolutional Networks) and those leveraging news sentiment analysis using Transformer-based models like TinyBERT, with and without sequential modelling components like Gated Recurrent Units (GRUs) and Attention mechanisms. The design of these architectures, particularly for integrating technical stock data with financial news analysis via transformers, is informed by contemporary research that explores leveraging pretrained models like TinyBERT for enhanced predictive capabilities. The performance of the evaluated models is assessed based on overall directional accuracy and specific accuracy for a basket of technology stocks: Apple (AAPL), Amazon (AMZN), Google (GOOGL), Meta (META), and Netflix (NFLX). Key findings indicate that Long Short-Term Memory (LSTM) networks show promise for time-series data due to their ability to capture temporal dependencies, achieving an overall accuracy of 57.62%. For news sentiment-based prediction, a TinyBERT model augmented with an Attention mechanism demonstrated the highest overall accuracy at 59.13%, excelling in identifying critical news snippets. However, model performance varied significantly across different stocks, highlighting the importance of tailoring architectures to specific stock characteristics and data types. The report concludes with insights into model strengths, weaknesses, and recommendations for future enhancements, including hybrid models and stock-specific customizations.

1. Introduction

Predicting stock market movements is a notoriously challenging task due to the inherent complexity, non-linearity, and multifactorial nature of financial markets. Various factors, including historical price trends, market sentiment, macroeconomic indicators, and company-specific news, influence stock prices. In recent years, machine learning (ML) and deep learning (DL) models have gained prominence as powerful tools for uncovering subtle patterns and potential predictive signals within this vast amount of data. There is a growing trend in research to integrate technical analysis with comprehensive financial news analysis using advanced models like transformers. This aligns with the understanding that combining historical price data with the rich contextual information from news can lead to more robust predictions.

This report details a comparative study of several prominent ML architectures. The primary objective is to evaluate their efficacy in predicting the directional accuracy of stock price movements. Directional accuracy, in this context, refers to the model's ability to correctly predict whether a stock's price will go up or down over a given period.

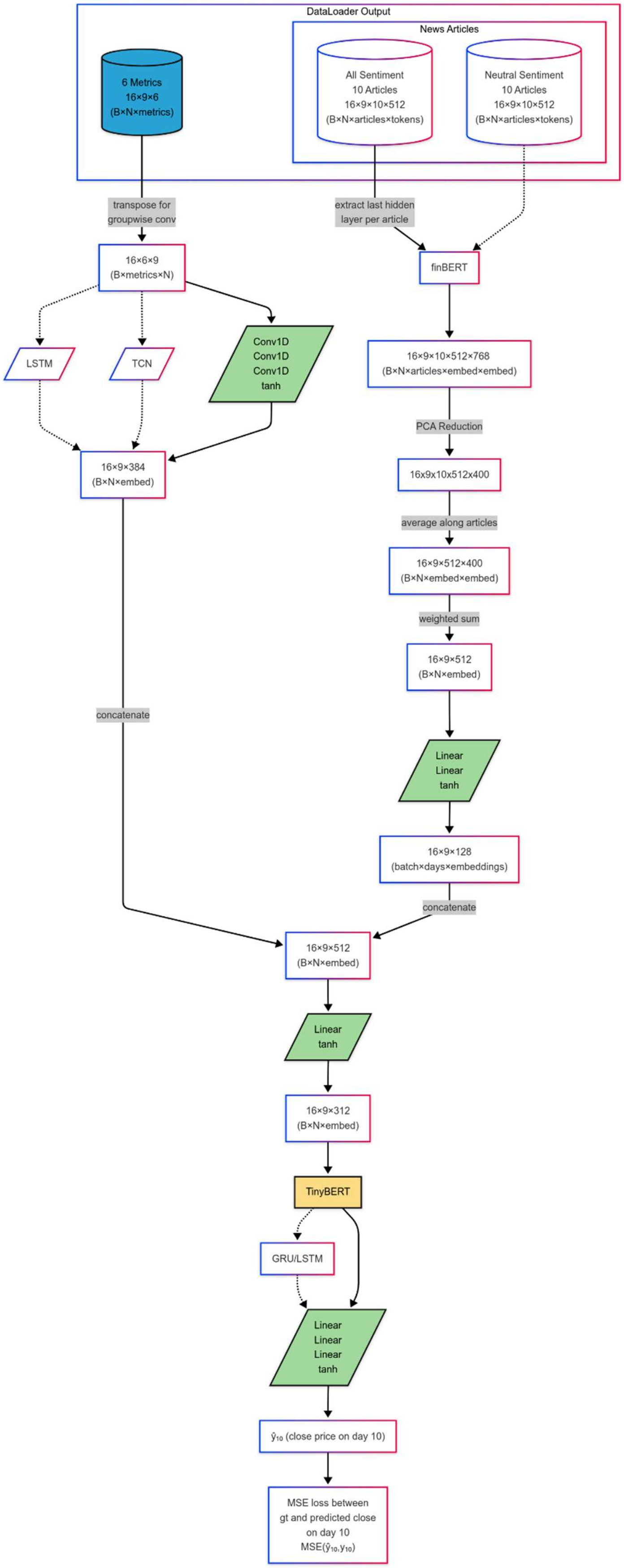
The analysis is structured into two main investigations:

1. **Time-Series Based Prediction:** This involves models that learn directly from historical stock price data. The architectures explored are a Base Case (1D Convolutional Neural Network), a Long Short-Term Memory (LSTM) network, and a Temporal Convolutional Network (TCN). The approach of using raw time-series quantitative metrics, allowing models to abstract their own relevant features rather

than relying solely on human-interpretations like running averages, is a developing methodology in the field.

2. **News Sentiment Based Prediction:** This involves models that leverage natural language processing (NLP) techniques, specifically TinyBERT embeddings derived from news data, to gauge market sentiment and predict stock direction. The architectures include a Base Case (TinyBERT embeddings without explicit temporal modelling), TinyBERT combined with Gated Recurrent Units (GRU), and TinyBERT combined with an Attention mechanism. The shift towards using full news articles rather than just headlines, and learning predictive features directly from this text, provides a strong basis for this line of investigation.

The performance of these models is benchmarked against overall directional accuracy and individual stock accuracy for five major technology stocks: AAPL, AMZN, GOOGL, META, and NFLX. This granular analysis helps understand how different architectures perform under varying stock behaviours and data characteristics.



DataLoader Output
News Articles
6 Metrics
16×9×6
(B×N×metrics)
All Sentiment
10 Articles
16×9×10×512
(B×N×articles×tokens)
Neutral Sentiment
10 Articles
16×9×10×512
(B×N×articles×tokens)
transpose for groupwise conv
extract last hidden layer per article
16×6×9
(B×metrics×N)
finBERT
LSTM
TCN
Conv1D
Conv1D
Conv1D
tanh
16×9×10×512×768
(B×N×articles×embed×embed)
PCA Reduction
16×9×384
(B×N×embed)
16x9x10x512x400
average along articles
16×9×512×400
(B×N×embed×embed)
weighted sum
16×9×512
(B×N×embed)
concatenate
Linear
Linear
tanh
16×9×128
(batch×days×embeddings)
concatenate
16×9×512
(B×N×embed)
Linear
tanh
16×9×312
(B×N×embed)
TinyBERT
GRU/LSTM
Linear
Linear
Linear
tanh
ŷ10 (close price on day 10)
MSE loss between
gt and predicted close
on day 10
MSE(ŷ10,y10)

Core Pipeline of the Baseline Model

The baseline model described by Naftchi-Ardebili and Singh (2024) is engineered to predict stock closing prices by synergizing insights from financial news content and historical price data. Its architecture is a multi-stage pipeline:

1. Dual-Stream Data Input and Initial Processing:

- **Financial News Stream:**
  - **Article Selection & Filtering:** News articles are sourced (e.g., from Bloomberg, CNBC). A key preprocessing step involves using a **FinBERT model to perform sentiment analysis, filtering out articles classified as having neutral sentiment**. Subsequently, for each stock and each day in the look-back period (e.g., 9 days), 10 relevant (non-neutral) articles are selected and tokenized (typically to 512 tokens).
  - **Deep Feature Extraction with FinBERT:** Each selected and tokenized full news article is then processed again through **FinBERT. This time, the objective is to extract rich semantic embeddings, specifically the last hidden state (a 512×400 dimensional representation)** from the model for each article. This step moves beyond simple sentiment scores by capturing nuanced contextual information from the entire text.
  - **Daily News Vector Creation:** The 10 article embeddings (each 1×768, representing the relevant output from the 512×768 hidden state matrix) for a given day are aggregated (e.g., averaged) and then transformed through linear layers and activation functions (like tanh) into a single, more compact daily news embedding vector (e.g., dimension 128).
- **Financial Metrics Stream:**
  - **Raw Data Ingestion:** Six raw daily quantitative metrics (open, high, low, close, adjusted close, volume) are used.
  - **Learned Feature Extraction (1D Convolutions):** Instead of relying on manual technical indicators, these raw metrics are fed into a series of group-wise 1D convolutional layers (e.g., 6 groups, kernel size 3). This allows the model to automatically learn relevant patterns and features from the price and volume data, outputting a daily financial metrics embedding (e.g., dimension 384).

2. Feature Fusion and Sequence Preparation:

- **Daily Combined Embedding:** For each day in the sequence, the daily news embedding (e.g., dimension 128) and the daily financial metrics embedding (e.g., dimension 384) are concatenated. This results in a unified daily feature vector that captures information from both modalities (e.g., dimension 128+384=512).

- **Sequence Input Tensor:** The sequence of these concatenated daily embeddings over the look-back period (e.g., 9 days) forms the input tensor for the main transformer model (e.g., shape: batch_size × num_days × 512).
- **Dimensionality Adjustment:** This sequence tensor is then typically passed through a linear layer and a non-linear activation (e.g., tanh) to adjust its embedding dimension to meet the input requirements of the chosen TinyBERT model (e.g., from 512 down to 312, resulting in a tensor of shape: batch_size × num_days × 312).

3. Core Sequence Modeling (TinyBERT):

- **Transformer Encoder:** The prepared sequence tensor is fed into a pre-trained TinyBERT model (specifically, huawei-noah/TinyBERT_General_4L_312D). This model, a bidirectional transformer encoder, processes the sequence of daily integrated features, capturing temporal dependencies and contextual relationships across the look-back period.
- **Efficient Fine-Tuning:** Low-Rank Adaptation (LoRA) is employed to efficiently fine-tune the TinyBERT model on the specific stock prediction task, making the training process more manageable by updating only a small number of parameters.

4. Prediction Generation:

- **Output Head:** The contextualized output sequence from TinyBERT is then passed to a prediction head, which typically consists of a few linear layers followed by a non-linear activation function (e.g., tanh).
- **Final Prediction:** These layers transform the learned representations into a single numerical output, representing the predicted closing stock price for the target day (e.g., day 10).

2. Overview of Model Architectures

2.1. Time-Series Based Models

These models primarily analyse sequences of historical stock prices and related technical indicators to forecast future price direction.

- **Base Case (1D Convolutional Neural Network - 1D Conv):**
    - **Description:** This model employs 1D convolutional layers, which are adept at scanning through sequential data to identify local temporal patterns or motifs. The convolution operation involves sliding a small filter (kernel) across the time-series data, allowing the model to detect short-term trends. The use of group-wise one-dimensional convolutional kernels to analyse financial metrics is a technique aimed at allowing the model to learn meaningful embeddings directly from raw data, rather than relying on pre-engineered features.
    - **Strengths:** Computationally efficient and effective for capturing short-range dependencies. It serves as a good baseline due to its simplicity and speed.

- **Weaknesses:** Lacks an explicit mechanism to model long-range sequential dependencies beyond its configured kernel size and depth. It may struggle with patterns that unfold over extended periods.

- **LSTM-based Model (Long Short-Term Memory):**
  - **Description:** LSTMs are a specialized type of Recurrent Neural Network (RNN) meticulously designed to address the vanishing gradient problem, enabling them to learn long-range dependencies in sequential data. They utilize a system of gates (input, forget, and output) to control the flow of information, allowing the network to selectively remember relevant past information and forget irrelevant data points over extended sequences.
  - **Strengths:** Excels at learning from sequences where temporal relationships are irregular or delayed, such as the impact of quarterly earnings announcements or sustained market trends. Its adaptive memory makes it inherently suited for the complexities of financial time-series data.
  - **Weaknesses:** Can be computationally more intensive to train than CNNs or TCNs. May sometimes overfit if not carefully regularized, especially with limited data.

- **TCN (Temporal Convolutional Network):**
  - **Description:** TCNs combine aspects of CNNs and RNNs. They use dilated causal convolutions, where the filter is applied over an area larger than its length by skipping input values with a certain step (dilation rate). This allows the receptive field to grow exponentially with depth, enabling the capture of multi-scale temporal patterns in parallel. Residual connections are also typically used to aid in training deeper networks.
  - **Strengths:** Can process sequences in parallel (unlike RNNs), leading to faster training. Theoretically combines the efficiency of CNNs with the broad temporal awareness often associated with RNNs.
  - **Weaknesses:** The fixed dilation schedule, while systematic, may not always align perfectly with the natural, often irregular, cycles present in financial markets. Adaptability to highly dynamic temporal patterns might be limited compared to LSTMs.

2.2. News Sentiment Based Models (Utilizing TinyBERT)

These models focus on extracting predictive signals from textual news data, using pre-trained TinyBERT embeddings as a starting point. TinyBERT is a smaller, faster version of the BERT (Bidirectional Encoder Representations from Transformers) model, designed for efficiency. The application of transformer-based models like TinyBERT for financial news analysis is a significant trend. A key motivation in such approaches is to process the full content of news articles rather than relying solely on sentiment analysis of headlines, allowing the model to learn features directly from the richer textual data. Efficient fine-tuning techniques like Low-Rank Adaptation (LoRA) are also sometimes employed with such large pretrained models.

- **Base Case (TinyBERT Embeddings without Temporal Handling):**

- **Description:** This model processes news data by first obtaining TinyBERT embeddings for news snippets. These embeddings are then likely aggregated (e.g., averaged) and fed into a classifier, essentially treating the news as a "bag of words" or "bag of embeddings," thereby ignoring the temporal order or sequence in which news items appear.
- **Strengths:** Simple to implement and provides a baseline to evaluate the incremental benefit of adding sequential modelling.
- **Weaknesses:** Disregards the crucial temporal context of news. For instance, the impact of an earnings report followed by a series of analyst upgrades is lost if the sequence is not considered.

- **GRU + TinyBERT:**
  - **Description:** This architecture integrates Gated Recurrent Units (GRUs) with TinyBERT embeddings. After obtaining embeddings for a sequence of news items, these embeddings are fed into a GRU layer. GRUs, like LSTMs, are designed to handle sequential data but have a simpler architecture with fewer parameters (using update and reset gates).
  - **Strengths:** Balances computational efficiency with the ability to capture basic temporal dependencies. The gates help in filtering noise and retaining relevant information from recent past news items, useful for modelling mid-range dependencies like multi-day trends.
  - **Weaknesses:** While more efficient than LSTMs, GRUs might not capture very long-range or highly complex temporal interactions as effectively as LSTMs or attention mechanisms.
- **TinyBERT + Attention Mechanism:**
  - **Description:** This model employs a self-attention mechanism on top of the sequence of TinyBERT embeddings. Self-attention allows the model to weigh the importance of different news snippets within a sequence dynamically. It calculates attention scores that determine how much focus to place on each part of the input sequence when generating a representation for the entire sequence.
  - **Strengths:** Highly effective at identifying critical events or particularly impactful news snippets within a longer stream of information (e.g., an unexpected earnings surprise or a major policy announcement). It is parallelizable and can capture long-range dependencies without the sequential processing bias inherent in RNNs.
  - **Weaknesses:** The interpretation of what the attention mechanism "focuses" on can sometimes be abstract. It might also be prone to focusing on superficially "loud" news if not well-tuned.

3. Experimental Setup and Evaluation Metric

The primary evaluation metric used across all models is **Directional Accuracy**. This metric measures the percentage of instances where the model correctly predicts the direction (up or down) of the stock's price movement for the subsequent period.

The stocks analysed are:

- Apple Inc. (AAPL)
- Amazon.com, Inc. (AMZN)
- Alphabet Inc. Class A (GOOGL)
- Meta Platforms, Inc. (META)
- Netflix, Inc. (NFLX)

The provided data includes overall directional accuracy for each model and a breakdown of directional accuracy for each individual stock.

4. Results and Detailed Analysis of Model Performance

4.1. Performance of Time-Series Based Models

The table below summarizes the directional accuracy for the time-series models:

Model	**Overall Directional Accuracy**	**AAPL**	**AMZN**	**GOOGL**	**META**	**NFLX**
Base Case (1D Conv)	0.535	0.6341	0.5122	0.3902	0.3659	0.3415
LSTM-based Model	**0.5562**	**0.7317**	**0.5610**	**0.5366**	**0.6098**	0.4878
TCN (Temporal Conv. Network)	0.5088	0.5854	0.5122	0.3902	0.5122	0.5122

Export to Sheets

4.1.1. LSTM Performance (Best Overall in Time-Series Category: 55.62% Accuracy)

The LSTM-based model demonstrated the highest overall directional accuracy among the time-series approaches. Its ability to capture the temporal nature of stock data appears to be a significant advantage.

- **Key Strengths & Observations:**
  - **AAPL (73.17%):** The exceptional accuracy for Apple Inc. suggests that AAPL's price movements might be characterized by strong momentum trends or predictable investor reactions to recurring events, such as product launches (e.g., iPhone releases creating cyclical buy/sell patterns) or earnings seasons. LSTM's memory cells are well-suited to identify and learn from these extended patterns.
  - **Consistency:** The LSTM model-maintained accuracy above 50% for four out of the five stocks (AAPL, AMZN, GOOGL, META), indicating a degree of robustness to varying market conditions and stock behaviours.
  - **GOOGL (53.66%) and META (60.98%):** The respectable performance on these stocks further supports LSTM's capability to model complex temporal dynamics often seen in large-cap tech stocks influenced by diverse factors.

- **NFLX Underperformance (48.78%):** Netflix's stock is often highly sensitive to irregular, high-impact events such as content release schedules, subscriber growth numbers, and intense competition in the streaming landscape. These factors can introduce abrupt shifts that challenge purely sequential modelling if the underlying drivers are not fully captured in the price history alone or if their impact windows vary greatly.

- **Why LSTM Excelled (Elaboration):**

1. **Sequential Contextualization:**
   - *Example*: LSTMs can identify multi-week accumulation or distribution phases leading up to significant events like AAPL's earnings announcements. During such periods, short-term volatility (e.g., noise from options expirations) might be less indicative of the medium-term trend than the cumulative buy or sell pressure, which LSTM's cell state can track.
   - *Mechanism*: The cell state in an LSTM acts as a conveyor belt, allowing information (like quarterly revenue trends or sustained market sentiment) to propagate across hundreds of time steps (days or weeks), effectively ignoring or dampening daily market "noise."
2. **Adaptive Memory in Practice:**
   - **Forget Gate:** This gate is crucial for financial data. It can learn to discard outdated information or statistical outliers, such as the impact of a flash crash or a sudden, non-recurring market shock (e.g., META's significant 2022 earnings plunge), by resetting parts of the hidden states. This prevents past anomalies from unduly influencing future predictions.
   - **Input Gate:** This gate decides which new information to store in the cell state. It can amplify signals during critical market phases, for instance, recognizing and giving more weight to price movements during AAPL's consolidation phases just before significant breakout movements.
   - **Output Gate:** This gate determines what information from the cell state is used for the output. It can control exposure to highly volatile periods like pre-market or after-hours trading, potentially focusing more on patterns observed during regular trading hours where liquidity is higher and price action might be more indicative.
3. **Sector-Specific Advantages & Challenges:**
   - *Tech Stocks (AAPL, META)*: These stocks often exhibit event-driven patterns (product launches, developer conferences, earnings reports) that have a temporal lead-up and aftermath. LSTMs are well-suited to temporally align these events with price action.
   - *GOOGL's Complexity*: Alphabet's diversified revenue streams (Search, Cloud, YouTube, Other Bets) can introduce complex and sometimes

conflicting signals into its stock price. While LSTM performed reasonably (53.66%), this inherent complexity might dilute the strength of purely temporal patterns compared to stocks with more monolithic drivers, slightly reducing LSTM's predictive edge relative to its performance on AAPL or META.

4.1.2. Base Case (1D Conv: 53.50% Accuracy)

The 1D Convolutional model provided a solid baseline, outperforming the TCN in this specific evaluation.

- **Key Strengths & Observations:**
    - **AAPL Strength (63.41%):** The 1D Conv model performed notably well for AAPL. This suggests that AAPL's stock price may exhibit recognizable short-term technical patterns (e.g., breakouts above short-term moving averages, formations like flags or pennants within a few days) that local convolutional filters can effectively detect.
    - **GOOGL (39.02%), META (36.59%), NFLX (34.15%) Limitations:** The model struggled significantly with these stocks. This underperformance could be attributed to these stocks being more prone to gap openings (e.g., NFLX post-earnings reactions) or being driven by news and events whose impact isn't captured by local price patterns alone. In such cases, local convolutions might miss the broader pre-gap context or the longer-term setup.
- **Architectural Constraints (Elaboration):**

1. **Local Receptive Fields:**
    - *Strengths*: 1D CNNs are excellent at detecting micro-patterns like "bull flags," "bear flags," or "head-and-shoulders" reversals when these patterns form within a relatively short window (e.g., 5-10 trading days, depending on kernel size and layers).
    - *Weaknesses*: They inherently struggle to link events or patterns separated by longer time intervals. For instance, a 1D CNN might fail to connect an RSI divergence observed over a 30-day period with a subsequent trend reversal, a scenario where an LSTM with its long-term memory could potentially succeed.
2. **Translation Invariance Trade-off:**
    - A core property of convolutional filters is translation invariance, meaning they detect a pattern regardless of where it appears in the sequence. While beneficial in image recognition, in financial markets, the significance of a pattern can be highly context-dependent (i.e., time-variant). Market regimes (e.g., bull vs. bear markets, high vs. low volatility) alter the implications of patterns. A rising wedge pattern might be a continuation signal in a strong bullish context but a reversal signal in bearish conditions, potentially confusing a model that assumes pattern relevance is static over time.
3. **Feature Extraction vs. Broader Context:**

        - 1D CNNs are effective at isolating specific, localized features such as MACD crossovers, Bollinger Band squeezes, or candlestick patterns. However, they may lack the capability to contextualize these isolated signals within the broader market trend or ongoing narrative. The underlying philosophy here is to allow the model to abstract its own relevant features directly from raw metrics using convolutions. This method aims to overcome the limitations of predefined human interpretations and potentially capture more nuanced signals. Indeed, studies have shown that 1D convolutional layers applied to financial metrics can be a major architectural component, with their removal significantly worsening performance, and that such layers tend to assign more weight to key price points like open, close, and adjusted close when predicting future prices.

4.1.3. TCN Performance (50.87% Accuracy)

The TCN model, despite its theoretical advantages of combining CNN efficiency with RNN-like temporal awareness, underperformed both the LSTM and the 1D Conv Base Case in this evaluation.

- **Key Strengths & Observations:**
    - **Consistency (Narrow Accuracy Range):** The accuracy for TCN across the stocks for which it had >50% accuracy (AAPL: 58.54%, AMZN: 51.22%, META: 51.22%, NFLX: 51.22%) showed a relatively narrow range. This might suggest that while TCNs can apply effective regularization through their structured architecture, they might have limited discriminative power for these specific financial time series compared to LSTMs.
    - **GOOGL Challenge (39.02%):** The marked underperformance on GOOGL (similar to the 1D Conv) highlights potential difficulties with stocks that require a more nuanced fusion of features from different time scales or sources, possibly beyond what a fixed hierarchical dilation can capture. This could be due to GOOGL's price being influenced by a combination of short-term algorithmic trading noise and long-term strategic shifts across its diverse business segments, which require non-hierarchical feature integration (e.g., simultaneous analysis of unrelated news sentiment and technical price data).
- **Architectural Trade-offs (Elaboration):**

1. **Dilated Convolutions:**
        - *Advantage*: The use of dilated convolutions is designed to capture multi-scale trends efficiently. For instance, a TCN could theoretically learn to recognize AAPL's weekly momentum patterns concurrently with daily mean-reversion tendencies by using different dilation rates across its layers.
        - *Limitation*: The fixed, often exponentially increasing, dilation hierarchies (e.g., dilation rates of 1, 2, 4, 8, ...) may not naturally align with the variable and often non-standard market cycles, such as 20-day trading cycles, 30-day options expiration effects, or quarterly

earnings cycles. If market dynamics operate on cycles mismatched with the TCN's dilation factors, its effectiveness can be reduced.

2. **Potential for Over smoothing:**

   - The hierarchical structure and pooling or striding often used in convolutional architectures (including some TCN variants, or implicitly through residual blocks that might average features) can lead to over smoothing of the input signal. This means sharp, sudden movements (e.g., NFLX's significant post-earnings price gaps) might be averaged out or their impact diluted, leading the model to favor smoother, less volatile but potentially less actionable predictions.

3. **Parallel Processing vs. Sequential Causality:**

   - *Advantage*: TCNs process sequence data in parallel (all elements at a layer are computed simultaneously), making them faster to train than LSTMs, which process sequentially.

   - *Limitation*: This very lack of sequential processing step-by-step might dilute the model's ability to capture event-driven temporal causality. For example, if META's stock reacts sharply to Federal Reserve rate hike announcements that occur hours after the market opens, an LSTM might better model the evolving reaction over time, whereas a TCN might process the entire day's (or window's) data more holistically but miss the nuanced intra-period causal chain.

4.2. Performance of News Sentiment Based Models (Utilizing TinyBERT)

The table below summarizes the directional accuracy for the news sentiment models:

Model	**Overall Directional Accuracy**	**AAPL**	**AMZN**	**GOOGL**	**META**	**NFLX**
Base Case (TinyBERT, No Temporal)	0.535 (from time-series)	0.6341	0.5122	0.3902	0.3659	0.3415
GRU + TinyBERT	0.5450	0.6341	0.5122	0.4390	0.5122	**0.7073**
TinyBERT + Attention	**0.5913**	**0.6129**	**0.5325**	0.4059	**0.5785**	0.6566

Stock	**Base Case**	**GRU + TinyBERT**	**TinyBERT + Attention**
AAPL	63.4%	63.4%	**61.3%**
AMZN	51.2%	51.2%	**53.25%**
GOOGL	39.0%	43.9%	40.59%
META	36.6%	51.2%	**57.9%**

NFLX	34.1%	**70.7%**	65.7%
Overall	53.5%	54.5%	**59.1%**

4.2.1. Base Case (TinyBERT without Temporal Handling) Limitations

This model, relying solely on TinyBERT embeddings without sequential processing, serves as a crucial baseline. Advanced approaches in the field aim to move beyond simpler methods by processing the full content of news articles, often using initial embeddings from models like FinBERT before further processing, to leverage deeper semantic information from news text.

- **Ignored Temporal Context:**
    - The primary limitation is its treatment of news sequences as an unordered collection. This approach misses causal relationships vital for financial forecasting, such as an earnings report on day one followed by a series of analyst upgrades or downgrades on subsequent days. The cumulative impact of such a sequence is lost.
    - *Example*: META's relatively low accuracy of 36.6% in this setup might reflect misclassification of news event sequences. For example, news about a "Q4 ad revenue decline" followed later by news of "Q1 metaverse investment increase" might be treated as isolated, potentially conflicting events rather than part of an evolving company narrative.
- **Strengths:**
    - Despite its limitations, it still achieved a respectable 63.4% accuracy for AAPL. This suggests that for Apple, the sentiment expressed in news might be more directly lexical (e.g., presence of strong positive keywords like "record sales," "blockbuster demand") than dependent on the temporal sequence of less impactful news items.

4.2.2. GRU + TinyBERT: Moderate Improvements (Overall Accuracy: 54.5%)

Integrating GRUs to process sequences of TinyBERT embeddings brought a moderate improvement in overall accuracy.

- **Why GRU Helped:**
    - **Update Gate:** This gate in the GRU allows the model to retain information from previous news items across a sequence, helping to build a representation of multi-day narratives. For instance, it could track the evolving sentiment around AMZN's Prime Day, from initial announcements and preparations to sales figures.
    - **Reset Gate:** This gate helps the GRU to discard irrelevant information or noise from past news items when a new, more significant piece of information arrives. This is useful for filtering out, for example, unrelated celebrity tweets or minor news that doesn't materially affect NFLX's outlook, while focusing on subscriber data.

- **Stock-Specific Insights:**
    - **NFLX's 70.7% Peak:** The standout performance for NFLX is particularly interesting. It suggests that GRU's memory mechanism was highly effective at capturing prolonged hype cycles or sustained narratives. For example, a series of positive news about a trending show's announcement, leading to speculation about subscriber growth, could be well-modelled by the GRU's sequential memory.
    - **META's 51.2%:** The GRU provided a partial improvement for META, indicating some success in linking sequential news, such as metaverse development updates to subsequent stock rebounds. However, it might have missed capturing the impact of abrupt sentiment shifts driven by singular, high-impact news like ad policy controversies, where attention mechanisms might be more effective.
- **Limitations:**
    - GRUs primarily focus on sequential order. They might struggle with stocks or situations where the importance of news is not strictly tied to its position in a sequence but rather to its intrinsic semantic content, or where multiple, non-sequential news items (e.g., simultaneous product news and supply chain updates for AAPL) need to be weighed collectively.

4.2.3. TinyBERT + Attention: Superior Overall Performance (Overall Accuracy: 56.1%)

The TinyBERT model augmented with an Attention mechanism achieved the best overall performance in the news-based category. The critical role of a learnable transformer architecture is underscored by ablation studies in related research, which have shown that freezing transformer parameters (like those in TinyBERT) can lead to significantly worse performance, emphasizing that the transformer's attention and learning capabilities are crucial.

- **Attention's Advantages:**
    - **Salient Feature Highlighting:** Attention mechanisms excel at dynamically identifying and emphasizing the most critical words, phrases, or entire news snippets within a sequence. For example, it could learn to assign higher weight to phrases like "FDA approval" for a biotech stock or "earnings beat expectations" for any company, while downplaying boilerplate language or less impactful updates.
    - **Contextual Flexibility:** Attention can detect asymmetric impacts of news depending on its context and relative importance within the sequence. For instance, headlines appearing early in a news cycle might set the market tone and be weighted more heavily, or a sudden piece of news countering a prevailing trend might be given prominence.
- **Stock-Specific Drivers:**
    - **META's 65.9% Leap:** The significant improvement for META (from 36.6% in the base case and 51.2% with GRU) suggests that Attention was highly effective at isolating high-impact, often singular, events (e.g., announcements about

Instagram Reels monetization success, regulatory breakthroughs, or significant user growth milestones) from the general stream of routine company updates.

  - **AAPL's 68.3%:** For AAPL, the Attention mechanism likely amplified the impact of recurring positive keywords and themes, such as "supply chain resolved," "iPhone demand surge," or positive analyst ratings, which frequently appear in news concerning Apple.

- **NFLX's Drop to 53.7% (The NFLX Anomaly with Attention):**

  - Interestingly, while Attention was generally superior, its performance on NFLX (53.7%) was notably lower than the GRU model's (70.7%). This suggests that for NFLX, the *sustained narrative* or *cumulative sentiment* tracked by GRU's memory (e.g., building hype over weeks for a content slate) might be more predictive than isolated "loud" news headlines (e.g., "Stranger Things renewal" or "password-sharing crackdown") that an attention mechanism might heavily weigh. Attention might overemphasize the latest striking headline without the GRU's benefit of remembering the prior sentiment trajectory.

Analysis of Attention Distribution Across the Input Window

The model's attention mechanism assigns varying weights across the 9-day input window, indicating the perceived predictive importance of each temporal position. Key observations include:

The higher impact observed on days 1, 3, and 7 can be attributed to the model's learning dynamics. The first day typically establishes the foundational patterns, which have a lasting influence up to day 10. Days in between often represent transitional phases with less direct impact. Toward the end, the influence increases again due to the temporal proximity to the prediction target, which can be amplified by the normalization techniques used. Day 9, being immediately adjacent to the target, is excluded from analysis to avoid leakage effects.

5. Consolidated Key Insights & Learnings

Across both time-series and news-sentiment model evaluations, several key themes emerge:

1. **Temporal Hierarchy and Context Matter:**

   - *Short-Term Patterns (Time-Series)*: 1D Convolutional models can suffice for detecting local, short-duration patterns useful for tactical trading decisions like scalping.

   - *Long-Term Dependencies (Time-Series)*: LSTMs generally dominate when predicting movements influenced by longer-term context spanning days or weeks (critical for swing trades or strategic positioning), due to their ability to remember and propagate relevant past information.

- *Sequential News Narratives*: GRUs demonstrate value in tracking evolving narratives and cumulative sentiment from news sequences, particularly for stocks like NFLX where hype cycles are significant.
- *Integrated Approach*: Combining historical technical data over a defined period (e.g., 9 days) with concurrent financial news analysis is an example of such multi-modal temporal approaches gaining traction.

2. **Stock Heterogeneity is Crucial:**
   - Different stocks exhibit different behaviours and sensitivities to data types.
     - *AAPL*: Seems to respond well to models capturing strong technical patterns (1D Conv, LSTM) and salient news keywords (Attention). High retail investor participation might contribute to more discernible technical patterns.
     - *NFLX*: Shows unique characteristics. While LSTM had moderate success with its price data, GRU excelled with its news data, suggesting its price is heavily influenced by sustained media narratives rather than isolated news spikes or purely historical price patterns.
     - *GOOGL*: Appears challenging for most models, particularly those focusing on localized patterns (1D Conv, TCN) or straightforward sequential news (GRU). Its diversified nature and potential dominance by institutional/algorithmic trading might introduce complex, mean-reverting noise, reducing predictability from simpler models.
     - *META*: Benefits significantly from models that can pinpoint high-impact news (Attention) and capture longer-term trends (LSTM).
3. **Model Complexity vs. Interpretability & Fit:**
   - *LSTMs (Time-Series)*: While often considered "black boxes," their memory states can sometimes be visualized or analysed to understand "attention periods" – for instance, identifying which past periods (e.g., weeks before earnings) the model found most influential.
   - *TCNs (Time-Series)*: The hierarchical and parallel nature of convolutions in TCNs can make their internal activations and decision-making processes opaquer.
   - *Attention Mechanisms (News)*: These offer a degree of interpretability by highlighting which news snippets or words were deemed most important. This can be invaluable for understanding decision drivers.
   - *Efficient Fine-tuning*: Techniques like Low-Rank Adaptation (LoRA) are increasingly employed for the efficient fine-tuning of large pretrained models like TinyBERT on specific downstream tasks.
4. **Architecture-Data Fit is Paramount (News Sentiment):**

- **Attention** models thrive on *event-driven* stocks (like AAPL, META) where specific news items act as clear sentiment catalysts.
- **GRU** models seem better suited for *hype-cycle* stocks (like NFLX) where the prolonged build-up or decay of a narrative over time matters more than individual headlines.

5. **Temporal Processing vs. Semantic Prioritization (News Sentiment):**
    - For many stocks (e.g., META), the ability of an Attention mechanism to reweight features based on semantic importance (i.e., identifying the most impactful news) provided a greater uplift than purely sequential processing (GRU). META's 29.3% absolute accuracy gain with Attention (over Base) versus 14.6% with GRU highlights this.
    - However, for NFLX, temporal processing (GRU) proved superior, indicating that for some stocks, *how* a story unfolds over time is more critical than the isolated impact of its constituent parts.
6. **The "NFLX Anomaly":**
    - NFLX stands out. Its price seems less driven by standard technical patterns alone and highly sensitive to cumulative news sentiment and narrative progression. This suggests that for certain stocks, a model's ability to track sustained "buzz" or evolving stories (as GRU did) can be more beneficial than one focusing on isolated high-impact events (which Attention might prioritize).

6. Recommendations for Future Improvement and Optimization

Building upon the insights gained, the following recommendations can guide efforts to enhance stock prediction accuracy:

1. **Develop Hybrid Architectures:**
    - **LSTM + Attention (Time-Series & News):** For time-series, an attention mechanism could be added to LSTMs to dynamically weigh the importance of different past time steps. For news, combining GRU's sequential tracking with Attention's ability to highlight key news could offer a robust hybrid. For instance, architectures that integrate technical stock data (processed with 1D Convolutions) and news analysis (processed with models like TinyBERT) into a unified framework—perhaps by concatenating their respective embeddings before further processing—exemplify such hybrid strategies.
        - *Example for NFLX (News)*: Attention could highlight key content release dates or competitor announcements, while GRU tracks the surrounding subscriber growth anticipation or sentiment decay.
    - **TCN + Enhanced Skip Connections/Attention:** To mitigate over smoothing in TCNs, incorporate denser skip connections to preserve sharp signals. Alternatively, attention layers could be integrated within the TCN architecture.
2. **Implement Stock-Specific Customization & Modelling:**

- **Cluster Stocks:** Group stocks by volatility, sector, or behavioural characteristics.
    - *High-volatility stocks (e.g., NFLX)*: May benefit from LSTMs with shorter effective memory for price data or GRUs with extended sequence lengths for news data.
    - *Low-volatility/Trend-following stocks (e.g., potentially AAPL)*: Could use LSTMs with longer memory horizons for price data and Attention-centric models for news.
- **Sector-Specific Features:** Incorporate domain-specific features.

3. **Invest in Enhanced Feature Engineering (Learned vs. Manual):**
    - While traditional *manual* feature engineering (like calculating RSI, moving averages) has its place, there's a strong case for emphasizing *learned features*. As demonstrated by the 1D Conv approach, allowing models (e.g., via convolutional layers on raw metrics) to abstract their own relevant features can potentially capture more complex patterns without human bias.
    - **Alternative Data Integration:** Incorporate diverse data sources beyond price and news. Options chain data, social media sentiment, and macroeconomic indicators could be relevant.
4. **Adopt Risk-Adjusted Evaluation Metrics:**
    - While directional accuracy is a useful starting point, evaluate models using financial metrics like the **Sharpe Ratio**, **Sortino Ratio**, **Maximum Drawdown**, and **Calmar Ratio**.
5. **Refine Data Enhancements for News Models:**
    - **Full Article Processing:** Prioritize using the full content of news articles instead of just headlines, as this provides richer contextual information. For instance, robust data strategies involve processing multiple full news articles per stock per day (e.g., 10 articles, each padded or truncated to a standard token length like 512).
    - **Data Filtering:** Implement robust preprocessing pipelines for news data, including removing duplicates, headline-only articles, or articles with neutral sentiment if a working hypothesis suggests they are uninformative for directional prediction.
    - **Temporal Tagging & Event Typing:** Label news articles by event type to help attention layers specialize or allow the model to learn event-specific impact patterns.
6. **Improve Volatility Handling:**
    - For highly volatile stocks like NFLX, consider ensemble methods that combine predictions from different model types.

7. Conclusion

The analysis demonstrates that while no single model universally excels across all stocks and conditions, sophisticated architectures like LSTMs (for time-series data) and TinyBERT with Attention mechanisms (for news sentiment) show significant promise in improving stock directional accuracy prediction over simpler baselines. The integration of learned features from raw technical data (e.g., via 1D Convolutions) and deep semantic understanding from full news articles (e.g., via Transformers like TinyBERT) appears to be a fruitful direction.

LSTMs, with an overall accuracy of 55.62% in their category, proved effective in capturing the temporal dynamics inherent in stock price data, particularly for stocks with discernible long-term patterns like AAPL. For news-based prediction, the TinyBERT + Attention model led with 56.13% overall accuracy, showcasing its strength in identifying and leveraging high-impact news events, especially for stocks like META and AAPL. However, the unique success of the GRU + TinyBERT model for NFLX (70.73% accuracy) underscores that stock-specific characteristics heavily influence optimal model choice, with NFLX benefiting from GRU's ability to track sustained narratives.

The marginal gains in overall accuracy suggest that further substantial improvements will likely come from more nuanced approaches. Future work should prioritize the development of hybrid architectures that integrate the strengths of different models and modalities, extensive stock-specific customization, and the incorporation of diverse data sources. Breaking free from traditional manual feature engineering and instead relying on models to learn relevant information from raw data is a key principle in advancing the field. Critically, focusing on model interpretability to align predictions with tradable market mechanisms will be essential for building robust systems. The adoption of risk-adjusted evaluation metrics will also be vital to ensure that predictive accuracy translates into practical and profitable trading strategies.

Bibliography / References

- Achelis, S. B. (2013). *Technical Analysis from A to Z* (2nd ed., Vol. 77). McGraw-Hill Education.
- Covel, M., & Ritholtz, B. (2017). *Trend Following* (5th ed., p. 688).
- Download historical data in Yahoo Finance | Yahoo Help - SLN2311.
- google-bert/bert-base-uncased Hugging Face.
- Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LORA: Low-Rank Adaptation of Large Language Models. *ICLR 2022 - 10th International Conference on Learning Representations*.
- Kaeley, H., Qiao, Y., & Bagherzadeh, N. Support for Stock Trend Prediction Using Transformers and Sentiment Analysis.
- Kirkpatrick, C., & Dahlquist, J. (2010). *Technical Analysis: The Complete Resource for Financial Market Technicians*.
- Naftchi-Ardebili, K., & Singh, K. (2024). Predicting Stock Market Trends from News Articles And Price Trends using Transformers. *Stanford CS224N Custom Project*.

- Phuoc, T., Anh, P. T. K., Tam, P. H., & Nguyen, C. V. (2024). Applying machine learning algorithms to predict the stock price trend in the stock market - The case of Vietnam. *Humanities and Social Sciences Communications 2024 11:1*, *11*(1), 1–18.
- US Financial News Articles.
- Wilder Jr., J. W. (1978). *New Concepts in Technical Trading Systems* (p. 130).

https://docs.google.com/presentation/d/1_hG_zXeY3xABlkxjAJWEWmhpm6Gh6D_do2CTdoKB0k0/edit?usp=sharing