Integrating Multi-Source Data with Sentiment Analysis and Language Models to Enhance Stock Market Decision Making

PhD Research Progress Seminar

Anandkumar Pardeshi
Registration No.: 28/01-03-2025
Under the Guidance of
Dr. Sujata Deshmukh

Department of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering, Bandra
(Permanently Affiliated to University of Mumbai)

January 03, 2026

Agenda

Expert Feedback: Key Focus Areas

Critical Analysis & Defense Preparation

- **Core Challenge**: Address complex data dynamics in market regimes
- **Methodology**: Justify every choice with "Why & How"
- **Defense**: Prepare counter-arguments with data/theory

Research Guidelines (Next Phase)

1. **Fundamental Need**: Benchmark static models during market shifts
2. **Framework**: Implement real-time dynamic fusion engine
3. **Validation**: Conduct rigorous ablation studies
4. **Temporal**: Experiment with lag structures & granularity
5. **Uncertainty**: Compare advanced quantification techniques

Goal: Transform feedback into robust research contributions

Stock Data Dimensions and Complexity

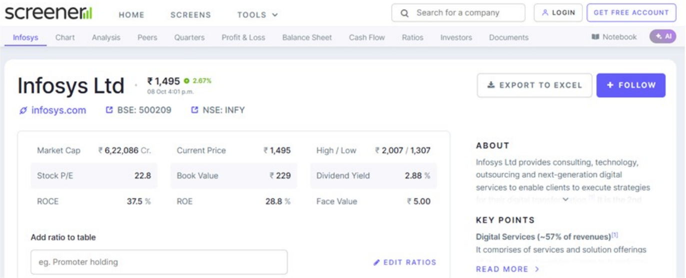
Stock data can be categorized into several dimensions, each adding complexity:

Type	Description	Example
Price Data	Historical prices over time	Open, High, Low, Close (OHLC)
Volume Data	Number of shares traded	Trade volume, order book
Fundamental Data	Company financials	EPS, P/E ratio, balance sheets
Sentiment Data	Public perception	News headlines, tweets
Macroeconomic Data	Market influencers	Interest rates, inflation, GDP
Alternative Data	Non-traditional	Satellite images, web traffic, ESG reports

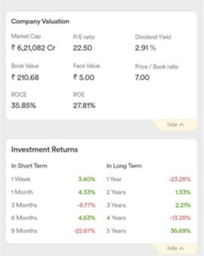
Each of these sources adds new layers of dimensionality and heterogeneity to the dataset.

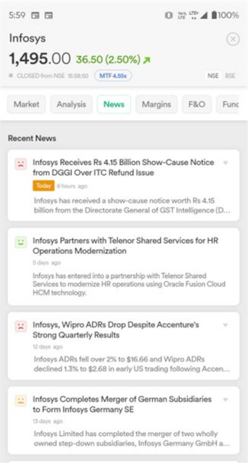
Infosys Ltd.

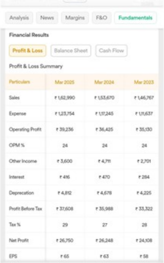

INFOSYS
Sell
Buy
Scalper
INFOSYS · 1D · NSE
Volume SMA 4.657M
EMA Cross 20 200
RSI 14 SMA 14
Jul
Sep
Nov
2025
Mar
May
Jul
Sep
Nov


screener
HOME
SCREENS
TOOLS
Search for a company
LOGIN
GET FREE ACCOUNT
Infosys
Chart
Analysis
Peers
Quarters
Profit & Loss
Balance Sheet
Cash Flow
Ratios
Investors
Documents
Notebook
AI
Infosys Ltd
₹ 1,495 2.67%
08 Oct 4:01 p.m.
EXPORT TO EXCEL
+ FOLLOW
infosys.com
BSE: 500209
NSE: INFY
Market Cap ₹ 6,22,086 Cr.
Current Price ₹ 1,495
High / Low ₹ 2,007 / 1,307
Stock P/E 22.8
Book Value ₹ 229
Dividend Yield 2.88 %
ROCE 37.5 %
ROE 28.8 %
Face Value ₹ 5.00
Add ratio to table
eg. Promoter holding
EDIT RATIOS
ABOUT
Infosys Ltd provides consulting, technology, outsourcing and next-generation digital services to enable clients to execute strategies
KEY POINTS
Digital Services (~57% of revenues)[1]
It comprises of services and solution offerings
READ MORE

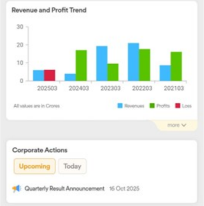
Infosys Ltd.







Particulars	Mar 2025	Mar 2024	Mar 2023
Sales	₹ 1,62,990	₹ 1,53,670	₹ 1,46,767
Expense	₹ 1,23,754	₹ 1,17,245	₹ 1,11,637
Operating Profit	₹ 39,236	₹ 36,425	₹ 35,130
OPM %	24	24	24
Other Income	₹ 3,600	₹ 4,711	₹ 2,701
Interest	₹ 416	₹ 470	₹ 284
Deprecation	₹ 4,812	₹ 4,678	₹ 4,225
Profit Before Tax	₹ 37,608	₹ 35,988	₹ 33,322
Tax %	29	27	28
Net Profit	₹ 26,750	₹ 26,248	₹ 24,108
EPS	₹ 65	₹ 63	₹ 58



India VIX - Volatility Index

What is VIX?

The India VIX measures the market's expectation of volatility over the next 30 days

HIGH VIX > 20

- **Fear & Uncertainty**
- Market falls likely
- High volatility
- **Time for Caution**

Fear Indicator

VS.

LOW VIX < 15

- **Complacency & Stability**
- Rising/steady markets
- Low volatility
- **Confidence high**

Complacency Indicator

VIX Level	Market Condition	Trading Implication
< 15	Bullish/Stable	Normal trading, trend following
15-20	Neutral/Transition	Cautious approach
> 20	Bearish/Volatile	Defensive strategies, hedging

FII/DII: Tracking Institutional Money Flow

Foreign Institutional Investors (FII)

Big money from abroad

Domestic Institutional Investors (DII)

Big money from within India

- Mutual Funds
- Insurance Companies
- Banks

Market Impact Scenarios

Scenario	Impact	Sentiment
FII BUY + DII BUY	Strong Uptrend	Very Bullish
FII SELL + DII SELL	Strong Downtrend	Very Bearish
FII BUY + DII SELL	Sideways/Volatile	Mixed
FII SELL + DII BUY	Sideways/Volatile	Mixed

Interpretation Guidelines

- **Same direction**: Strong trend confirmation
- **Opposite direction**: Market indecision/volatility
- **FII flows**: Often drive major trends
- **DII flows**: Provide domestic support

The Core Challenge: Complex Multi-Source Data

Problem Statement

Stock market prediction requires integrating heterogeneous data sources:

- **Numerical time-series** (OHLCV, technical indicators)
- **Unstructured text** (news, social media)
- **Alternative data** (India VIX, satellite imagery)

Current Limitations

- **Static fusion models** fail to adapt to market regime shifts.
- **Noisy sentiment data** due to sarcasm, ambiguity, financial nuance.
- **Lack of uncertainty-aware weighting** in multi-source integration.

Basic Statistics Overview

Publication Database Analysis

Dataset Summary

- **Total Publications Analyzed:** 51
- **Time Range:** 2020 - 2026 (7 years)
- **Unique Journals/Conferences:** 26

Yearly Distribution

Year	2020	2021	2022	2023	2024	2025	2026
Count	1	4	1	9	26	9	1

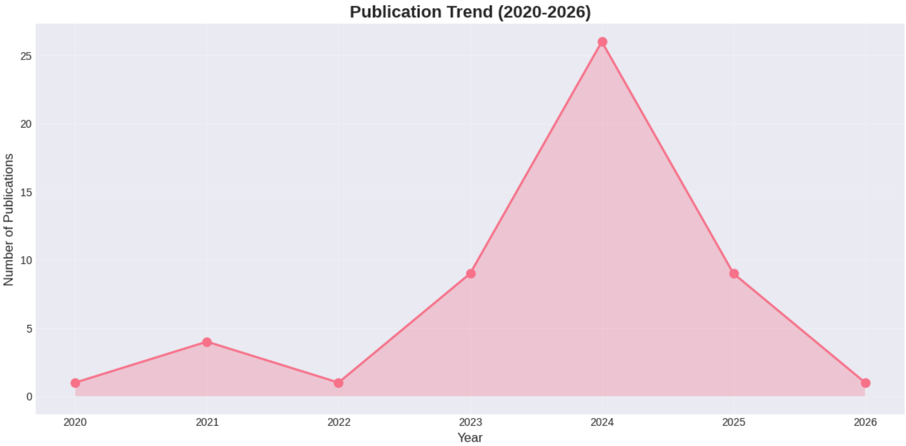
Key Observation

2024: Peak Publication Year
26 publications (51% of total)
Explosive growth in LLM research

Publication Density

- Avg: 7.3 publications/year
- 2023-2025: 86% of publications
- Clear research acceleration

Publication Trend



Top 10 Journals/Conferences

Publication Venues Distribution

Journal/Conference	Count	Percentage
arXiv	13	25.5%
IEEE Access	12	23.5%
Information Fusion	2	3.9%
Springer Nature Singapore	2	3.9%
Applied Artificial Intelligence	1	2.0%
MATEC Web Conf.	1	2.0%
AI	1	2.0%
Heliyon	1	2.0%
Investment Management and Financial Innovations	1	2.0%
SN Bus Econ	1	2.0%

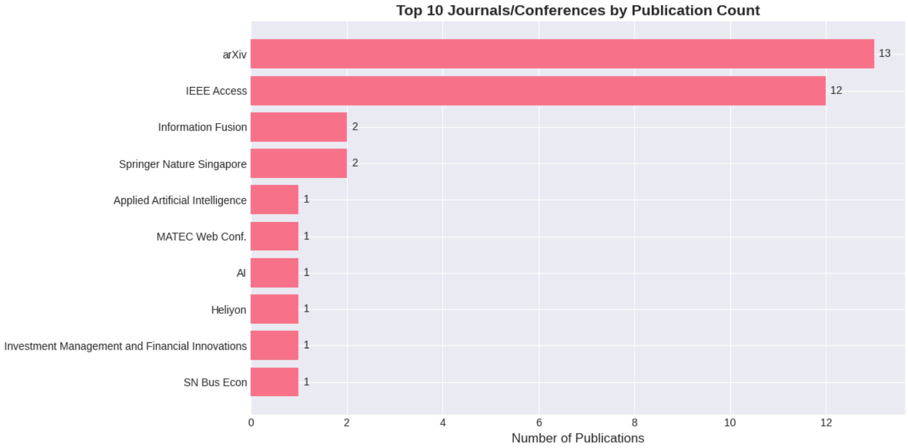
Key Insight

- **arXiv:** Leading venue (13 papers)
- **IEEE Access:** Close second (12 papers)
- Together: **49%** of publications

Publication Trends

- Pre-print servers dominant
- IEEE remains influential
- Diversified but concentrated

Top 10 Journals by Publication count



Keyword Analysis from Titles

Top 20 Keywords Identified

Top 10 Keywords

Keyword	Frequency
language	19
large	17
sentiment	17
learning	10
forecasting	10
review	6
enhancing	6
multi	5
machine	5
techniques	5

Research Focus Areas

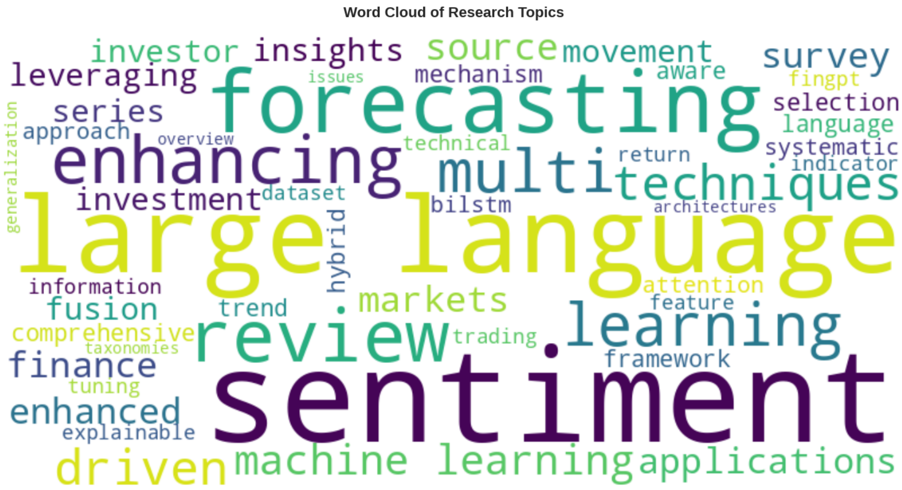
Primary Themes:

1. **Language Models** (36 mentions)
2. **Sentiment Analysis** (17 mentions)
3. **Forecasting** (10 mentions)
4. **Learning Methods** (10 mentions)

Key Insight

"Large Language" appears 36 times
Indicates strong LLM research focus

Word Cloud of Research Topics



Research Focus Areas Analysis

Categorization of Publications by Research Theme

Research Category	Count	Percentage
Forecasting	29	56.9%
LLM/Foundation Models	25	49.0%
Sentiment Analysis	18	35.3%
Survey/Review	12	23.5%
Deep Learning	9	17.6%
Multi-source Fusion	6	11.8%

Dominant Research Areas

Top 3 Categories:

1. **Forecasting** (56.9%)
2. **LLM/Foundation Models** (49.0%)
3. **Sentiment Analysis** (35.3%)

Overlap Analysis

- Publications often span multiple categories
- LLM + Forecasting: Common combination
- Sentiment + LLM: Growing synergy
- Multi-source + DL: Emerging trend

Research Trends and Observations

Emerging Patterns in LLM and Financial Research

Observed Trends

1. **LLM/Foundation Models research** is rapidly growing
2. **Multi-source fusion approaches** are gaining popularity
3. **Sentiment analysis remains** a key focus area
4. **Increased emphasis on** explainability and hybrid models

Publication Concentration

- **arXiv + IEEE Access:** 49% of publications
- **2024 alone:** 51% of total output
- **Top 3 categories:** Cover 79% of research

Growth Pattern

Three-Phase Evolution:

- 2020-2022: Foundation (6 papers)
- 2023-2024: Explosion (35 papers)
- 2025-2026: Specialization (10 papers)

Focus

Research focus shifting from traditional ML to LLM-driven approaches

Financial-Specific Language Models

Table: LLM-Based Models for Financial Applications
(Focus: Open-source/Foundation Models)

Publication Year	Primary Authors	Paper Title	Core Technology or Model	Methodology Focus	Key Financial Application
2024	Y. Liang et al.; H. Yang, X.-Y. Liu, and C. D. Wang	FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs; FinGPT: Open-Source Financial Large Language Models	FinGPT	Sentiment analysis and Data fusion	Stock movement prediction and financial analysis
2024	D. Mai	StockGPT: A GenAI Model for Stock Prediction and Trading	StockGPT	Time-series forecasting	Trend forecasting
2023	S. Wu et al.	BloombergGPT: A Large Language Model for Finance	BloombergGPT	Natural Language Processing	Financial analysis

General LLM-Based Approaches for Stock Prediction

Table: General LLM-Based Approaches for Stock Prediction
(Focus: Generic LLMs without specific model names)

Publication Year	Primary Authors	Paper Title	Core Technology or Model	Methodology Focus	Key Financial Application
2024	Z. Zhao and R. E. Welsch	Aligning LLMs with Human Instructions and Stock Market Feedback in Financial Sentiment Analysis	LLM	Sentiment analysis	Movement prediction
2025	R. Wang, M. Sun, and L. Wang	From news to trends: a financial time series forecasting framework with LLM-driven news sentiment analysis and selective state spaces	LLM	Time-series forecasting	Trend forecasting
2025	L. Alson Mantshimuli and J. Weirstrass Muteba Mwamba	Enhancing portfolio optimization with multi-LLM sentiment aggregation: A Black-Litterman integration approach	Multi-LLM	Sentiment analysis	Portfolio optimization
2024	L. J. Kurisinkel, P. Mishra, and Y. Zhang	Text2TimeSeries: Enhancing Financial Forecasting through Time Series Prediction Updates with Event-Driven Insights from Large Language Models	LLM	Time-series forecasting	Trend forecasting
2023	Y. Ding et al.	Integrating Stock Features and Global Information via Large Language Models for Enhanced Stock Return Prediction	LLM	Data fusion	Stock return prediction
2026	H. Phalangpatanakij et al.	Stock Price Prediction Using Univariate and Multivariate Historical Data with Post-Interpretation via Large Language Models	LLM	Data fusion	Stock price prediction

BERT-Based and Hybrid LLM Approaches

Table: BERT-Based and Hybrid LLM Approaches
(Focus: BERT variants and hybrid LLMs)

Publication Year	Primary Authors	Paper Title	Core Technology or Model	Methodology Focus	Key Financial Application
2024	O. Shobayo et al.	Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic Regression: A Data-Driven Approach	FinBERT and GPT-4	Sentiment analysis	Stock price prediction
2024	E. Zhu and J. Yen	BERTopic-Driven Stock Market Predictions: Unraveling Sentiment Insights	BERTopic	Sentiment analysis	Stock market prediction

Table: Deep Learning Architectures (Non-LLM Focus)
(Focus: CNN, LSTM, Attention-based models)

Publication Year	Primary Authors	Paper Title	Core Technology or Model	Methodology Focus	Key Financial Application
2024	K. Xu and B. Purkayastha	Enhancing Stock Price Prediction through Attention-BiLSTM and Investor Sentiment Analysis	Attention-BiLSTM	Sentiment analysis	Stock price prediction
2024	A. Luo et al.	Short-Term Stock Correlation Forecasting Based on CNN-BiLSTM Enhanced by Attention Mechanism	CNN-BiLSTM	Time-series forecasting	Correlation forecasting
2024	S. Latif et al.	Enhanced prediction of stock markets using a novel deep learning model PLSTM-TAL in urbanized smart cities	PLSTM-TAL	Time-series forecasting	Stock market prediction

Research Objectives

Objective 1: To study and create a model that seamlessly combines various data sources—such as financial indicators, news, social media sentiment, and macroeconomic indicators—for a comprehensive analysis that enhances stock prediction accuracy.

Objective 2: To develop a hybrid model that integrates large language models (LLMs), sentiment analysis, and traditional financial indicators, ensuring the model's scalability across different stock markets and optimizing deep learning architectures for real-time stock prediction.

Objective 3: To create and analyze an explainable AI framework that provides interpretable predictions for stock selection, studying the impact of noise in sentiment data and investigating the temporal aspects of sentiment data in relation to stock price movements.

Objective 4: To evaluate and validate the scalability and practical application of the proposed model across different stock markets, exploring transfer learning techniques for cross-market applications and validating the model against benchmark datasets and real-world market data.

Fundamental Need: The Problem Identified

Current multi-source financial forecasting models use static fusion methods (fixed weights, simple averaging, or static attention mechanisms). These methods make incorrect assumptions:

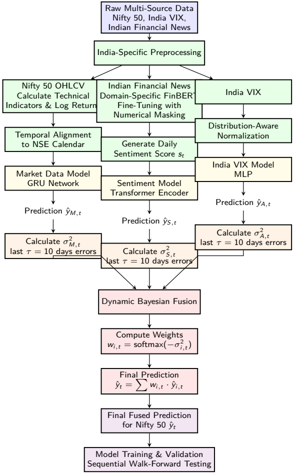
- Data source reliability is constant over time
- All sources are equally credible at all times

Why This Assumption Fails: Financial markets are non-stationary and experience regime shifts:

- In stable bull markets → Technical indicators work well
- During news-driven crashes → Sentiment data becomes crucial
- In high-volatility periods → Volatility indices (like India VIX) become important

Simple Example: If a model always gives 70% weight to technical indicators and 30% to sentiment, it will fail when news breaks that causes a market crash (when sentiment should get 90% weight instead).

Proposed Dynamic Fusion Framework for Multi-Source Financial Data



Data Preparation: Gather and clean three key Indian market data sources—Nifty price history, financial news, and the India VIX fear index—aligning them to the National Stock Exchange calendar.

Specialized Model Training: Each data type is processed by a dedicated AI model: a GRU network for market data, a FinBERT transformer for news sentiment, and an MLP for volatility data, generating independent predictions.

Dynamic Reliability Scoring: Continuously measure the recent prediction error (uncertainty) of each model over a rolling 10-day window to determine its current trustworthiness.

Intelligent Prediction Fusion: A Bayesian fusion mechanism dynamically assigns higher weight to the more reliable models, creating a single, robust final prediction that adapts to changing market conditions.

Validation: The entire system is validated using Sequential Walk-Forward Testing, simulating real-world trading to ensure performance is genuine and not based on historical bias.

DYNAMIC FUSION FRAMEWORK

Core Idea: A system that automatically adjusts trust in data sources based on recent performance.

Three-Layer Architecture:

RAW DATA → SPECIALIZED MODELS → UNCERTAINTY CALCULATOR → DYNAMIC WEIGHTS → FINAL PREDICTION

Step-by-Step Process:

1. **Data Collection:**
   - **S1:** Nifty 50 technical data — **S2:** Financial news sentiment — **S3:** India VIX
2. **Expert Model Training:**
   - **GRU** (Technical) — **Transformer** (Sentiment) — **MLP** (Volatility)
3. **Fusion Logic:**
   - Calculate daily uncertainty → Convert to dynamic weights → Combine predictions

CONTRIBUTION & NOVELTY

Key Innovations

- **Dynamic Weighting:** Fusion weights change daily based on real-time performance.
- **Uncertainty-Driven:** Weighting is based on *prediction confidence*, not just historical accuracy.
- **Bayesian Framework:** A formal probabilistic approach to weight calculation.
- **Self-Correcting:** Automatically down-weights unreliable sources without manual intervention.

Method Comparison

Method	Nature
Early Fusion	Static
Late Fusion	Static
Attention Models	Static
This Study	**Dynamic**

"Unlike existing models that learn fixed patterns, our weights adapt daily to market shifts."

Mathematical Model

Core Formula - Dynamic Weight Calculation:

$$w_{i,t} = \frac{\exp(-\sigma^2_{i,t})}{\sum_j \exp(-\sigma^2_{j,t})}$$

Where:
$w_{i,t}$ = weight for source i at time t
$\sigma^2_{i,t}$ = uncertainty/variance of source i at time t
exp() = exponential function (small uncertainties get big weights)

Uncertainty Calculation:

$$\sigma^2_{i,t} = \frac{1}{\tau} \times \sum_{k=1}^{\tau} (\text{error}_{i,t-k})^2$$

Where: $\tau = 10$ days (lookback window)

Final Prediction:

$$\hat{y}_t = w_{M,t} \cdot \hat{y}_{M,t} + w_{S,t} \cdot \hat{y}_{S,t} + w_{A,t} \cdot \hat{y}_{A,t}$$

Example: Stable Market

Day 1-10: Stable Market

- **Technical model error:** [0.1, 0.2, 0.1, 0.3, 0.2, 0.1, 0.2, 0.1, 0.3, 0.2] $\rightarrow \sigma_M^2 = 0.0076$
- **Sentiment model error:** [0.3, 0.4, 0.5, 0.3, 0.4, 0.6, 0.5, 0.4, 0.3, 0.5] $\rightarrow \sigma_S^2 = 0.011$
- **VIX model error:** [0.4, 0.5, 0.6, 0.4, 0.5, 0.7, 0.6, 0.5, 0.4, 0.6] $\rightarrow \sigma_A^2 = 0.017$

Initial Weights Calculation:

- $w_M = \exp(-0.0076) = 0.9924$
- $w_S = \exp(-0.011) = 0.9891$
- $w_A = \exp(-0.017) = 0.9831$
- **Sum** $= 0.9924 + 0.9891 + 0.9831 = 2.9646$

Final Weights: $w_M = 0.335$ (33.5%), $w_S = 0.334$ (33.4%), $w_A = 0.331$ (33.1%)
$\rightarrow$ *Technical gets slightly higher weight.*

Example: Market Crash

Day 11: Market Crash (News-Driven)

- **Technical model error spikes:** 0.8 (big miss)
- **Sentiment model error:** 0.2 (good prediction)
- **VIX model error:** 0.3 (moderate)

New uncertainties (using last 10 days including the crash):

- σ_M^2 increases to 0.05 (much higher)
- σ_S^2 decreases to 0.008 (lower)
- σ_A^2 increases slightly to 0.02

New Weights Result:

- w_M **decreases sharply** (maybe to 20%)
- w_S **increases** (maybe to 50%)
- w_A **moderate** (30%)

$\rightarrow$ **Model automatically trusts sentiment more during the crash.**

Interdependence: Dynamic Weight Redistribution

Key Insight: Sources aren't independent—poor performance in one automatically shifts the trust balance to others.

How It Works:

- **Scenario:** If the Technical model becomes unreliable ($\sigma_M^2 \uparrow$)
  - Its specific weight decreases ($w_M \downarrow$)
  - The available weight is automatically redistributed to Sentiment (S) and Volatility (A).
- **Equilibrium:** If S and A also become unreliable, the system forces all weights to become more equal to prevent over-reliance on a single weak source.

Failure Mode Protection

- **Black Swan Detection:** The system monitors if *ALL* sources become unreliable simultaneously.
- **Conservative Strategy:** During events where nothing works well, the model automatically defaults to equal weights to minimize the risk of extreme error.

Threshold : Temperature Parameter (T)

Modified Formula with Temperature Control:

$$w_{i,t} = \frac{\exp(-\sigma^2_{i,t}/T)}{\sum_j \exp(-\sigma^2_{j,t}/T)}$$

Purpose of T: Controls how aggressively weights respond to uncertainty:

- **Low T (e.g., 0.1): Aggressive** – Small uncertainty differences cause massive shifts in weights.
- **High T (e.g., 10): Conservative** – Weights change slowly; the distribution remains more stable.
- **Optimal T:** Usually found through validation (this study likely uses $T = 1$ as default).

Why It Matters

- **Noise Filtering:** Prevents over-reaction to temporary market noise.
- **Stability:** Ensures weights do not fluctuate wildly between trading days.
- **Customization:** Can be tuned for different institutional risk appetites.

Publications

1. A. Pardeshi and S. Deshmukh, "Deep Learning in Stock Market Forecasting: Comparative Insights and Future Directions," in *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)*, Navi Mumbai, India: IEEE, June 2025, pp. 1–6. doi: 10.1109/ICETI4T63625.2025.11132153.
[Scopus Indexed]
2. A. Pardeshi and S. Deshmukh, "AI in Finance: Computational Methods for Market Analysis and Risk Management," in *Next-Generation Computational Intelligence: Trends and Technologies*, vol. 60, S. Mahajan and J. B. De Vasconcelos, Eds., in Information Systems Engineering and Management, vol. 60., Cham: Springer Nature Switzerland, 2025, pp. 137–164. doi: 10.1007/978-3-031-96871-6_6.
[Indexed by Google Scholar. All books published in the series are submitted for consideration in the Web of Science.]

Thank You

Your insights would help improve this research—
please share any suggestions!

Anandkumar Pardeshi
PhD Candidate, Computer Engineering
Fr. Conceicao Rodrigues College of Engineering, Bandra