SVKM'S NARSEE MONJEE INSTITUTE OF MANAGEMENT STUDIES
(NMIMS)
NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS & ANALYTICS
DEPARTMENT OF STATISTICS

Capstone Report - April 2025

Sentiment Meets Sequence: A Deep Learning Pipeline for Real-Time Stock Price Fluctuation Prediction Using News and Historical Stock Data

Jerin Mathew (86032200035),
Jia Bindra(86032200028)

supervised by
Dr Revathy M,
Assistant Professor

Preface

In the modern age of digital finance, the stock market has emerged as a barometer of economic sentiment, reflecting not only tangible metrics such as revenue and profit but also the intangible tides of investor psychology, geopolitical shifts, and market speculation. With the exponential growth of machine learning and artificial intelligence, we are now empowered to decode patterns that were once obscured by complexity and scale.

This project is the culmination of months of research, experimentation, and analysis. It reflects our pursuit to build a predictive system that goes beyond conventional methods — one that draws from the nuanced world of financial news headlines and fuses it seamlessly with structured time-series data of stock prices. By leveraging state-of-the-art NLP models like DistilBERT and BART, in conjunction with LSTM networks for sequence modeling, we aimed to capture the multifaceted drivers of market movement.

The core idea germinated from a curiosity: Can the pulse of market sentiment, encapsulated in the brevity of a headline, forecast the future of a stock's price? This question led us into an in-depth exploration of multi-modal learning techniques, sentiment embedding strategies, and the peculiar intricacies of the Bombay Stock Exchange (BSE). The choice to focus on 53 actively traded companies from the 2022–2023 window gave us a data-rich yet temporally bounded domain within which to experiment and validate our approach.

This report stands as a comprehensive account of our journey — from data collection and preprocessing to model deployment and interpretation — and serves both as an academic artifact and a technical blueprint for future financial prediction systems.

Acknowledgements

We would like to express our profound gratitude to all those who have extended their support and guidance throughout the course of this project.

At the outset, we are deeply honored to acknowledge **Dr. Narayani Ramachandran**, Director, for her exemplary leadership, continuous encouragement, and for providing an environment conducive to academic excellence. Her vision and support have been instrumental in enabling us to undertake and complete this project.

We are sincerely grateful to **Dr. Revathy M, Dr. Anwesha Chattopadhyay**, and **Dr. Kavya S** for their invaluable guidance, constructive feedback, and consistent support during the various stages of this research. Their expertise and insights have greatly contributed to the depth and quality of our work.

We would also like to extend our appreciation to **Dr. Dileep Menon** for his technical assistance and thoughtful suggestions, which significantly enhanced the scope and analytical rigor of our project.

Finally, we express our heartfelt gratitude to our parents for their unwavering support, encouragement, and understanding throughout this academic journey. Their constant motivation has been a source of strength. We acknowledge with thanks all individuals and institutions who have contributed, directly or indirectly, to the successful completion of this project.

CERTIFICATE

I certify that the capstone project on **Temporal Sequence Transformer(TST)** submitted by **Jerin Mathew and Jia Bindra** in partial fulfillment of the degree of Bachelor of Science in Data Science,SVKM's,NMIMS(Deemed to be University), Mumbai,India is a original work carried out under my guidance.

Dr. M. Revathy
(Project Mentor)
Assistant Professor
School of Commerce
NMIMS, Bengaluru

Dr. Anwesha Chattopadhyay
Program Chair
Assistant Professor
SOMASA
NMIMS, Bengaluru

External

Dr. Narayani Ramachandran
Director
NMIMS, Bengaluru

Abstract

In the rapidly evolving landscape of technology and algorithmic trading, the demand for accurate, robust, and real-time predictive systems is greater than ever. This project addresses the challenge of forecasting short-term stock price changes by leveraging both structured and unstructured data sources. Through the integration of historical stock data and financial news sentiment embeddings, we propose a multi-modal deep learning framework that utilizes advanced techniques in natural language processing (NLP) and time series modeling. The model is specifically designed to predict the percentage change in stock prices for the next trading day. The core of our architecture combines DistilBERT-based sentiment embeddings and Long Short-Term Memory (LSTM) layers, enabling the system to interpret contextual information from news and temporal patterns in stock movement. With data drawn from the top 53 actively traded companies on the Bombay Stock Exchange (BSE) between 2022 and 2023, the system shows promising performance across several industry sectors. This approach demonstrates how hybrid AI models can enhance forecasting, bridging the gap between qualitative sentiment and quantitative pricing data.

Keywords

Multi-Modal Learning, DistilBERT Embeddings, LSTM (Long Short Term Memory) Networks, Stock Price Prediction, News Impact Modeling

Contents

1 Introduction

1.1 Background

The stock market is a complex, multifaceted system influenced by a wide range of factors including corporate performance, global economic events, public sentiment, and macroeconomic indicators. Traditionally, predictive models in finance have heavily relied on structured numerical data such as historical stock prices, volumes, and technical indicators. However, such models often overlook the immense influence of market sentiment, which can drive investor behavior and, consequently, price fluctuations. In recent years, advancements in natural language processing (NLP) have enabled deeper insights into unstructured textual data such as financial news and social media commentary. These developments open new opportunities for integrating sentiment into stock prediction models, allowing more comprehensive and responsive forecasting systems.

1.2 Problem Statement

Despite the growth of machine learning models in stock price prediction, most frameworks fail to incorporate the rich information embedded in textual data such as financial news. There is a critical need to develop a model that effectively combines both structured numerical data and unstructured sentiment information to better anticipate short-term price fluctuations. The key question guiding this research is, can a hybrid deep learning model that integrates BERT-based sentiment analysis with time series modeling accurately predict the percentage change in stock prices across diverse sectors on a day-to-day basis?

1.3 Objectives of the Project

Primary Objective:

To design and implement a deep learning model that forecasts the percentage change in next-day stock prices by fusing sentiment representations and historical stock data.

Secondary Objectives:

• Utilize fine tuned DistilBERT model to extract meaningful sentiment embeddings from financial news headlines.

• Apply BART-based sector classification to improve tagging accuracy and content

tion Technology, Services, Telecommunication, and Utilities. Initial exploration involved identifying the top 100 active companies in BSE as of 2025. Through rigorous analysis and data verification, it was determined that only 53 of these companies had consistent and reliable historical data available for the 2022–2023 timeframe. These 53 companies form the backbone of our modeling efforts, ensuring data relevance and integrity.

Furthermore, while the current model focuses on a single-day lookback, future iterations aim to expand this temporal window to capture more complex, longer-term dependencies. These trading gaps were carefully handled to maintain alignment between news sentiment data and actual trading days, thereby preventing data leakage and improving model accuracy.

2 Literature Review

The journey toward reliable and intelligent stock market prediction has evolved from classical econometric models to modern AI-powered, sentiment-driven forecasting systems. One of the foundational steps in this evolution is the FinBERT-LSTM model, which integrates FinBERT, a transformer-based model fine-tuned on financial texts—to generate sentiment embeddings, and feeds them along with historical price data into a stacked LSTM network for time-series forecasting. While powerful, the model proposed by the unnamed authors of this approach suffers from several key limitations: it applies static sentiment scores without contextual adaptability, lacks ticker and sector specificity, and assumes uniform market behavior across diverse financial entities.

In parallel, researchers such as **Narayana Darapaneni et al.** explored hybrid approaches tailored for the Indian stock market. Their study, **"Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets"**, introduced a dual-model strategy—using LSTM for historical price modeling and a Random Forest regressor powered by sentiment scores and macroeconomic indicators like gold prices, oil, and bond yields . Their use of sentiment from curated financial news aligned with the goals of FinBERT-LSTM but highlighted the gap in automated, real-time sentiment extraction from Indian sources.

Another regional perspective came from **Madhusmita Khuntia and Deepa Gupta**, who proposed a headline classification model using BERT-based embeddings and LSTM architectures to classify Indian news headlines. Their work, though primarily aimed at classification rather than regression, underlined the power of domain-adapted embeddings and sequence models in dealing with short, context-rich text like financial headlines .

From a higher vantage point, **Rahul Jain and Rakesh Vanzara** provided a systematic overview in their paper **"Emerging Trends in AI-Based Stock Market Prediction"**. They emphasized the growing role of LSTM, CNN, and reinforcement learning in modern prediction pipelines and drew attention to the importance of feature selection and non-linear modeling for decoding complex financial dynamics . Their survey revealed that while sentiment analysis and LSTM networks dominate the current research landscape, few systems account for domain adaptability or regional financial news, as is crucial for Indian market applications.

Adding another layer of sophistication, **Pranav Putta et al.** introduced Agent Q, a reasoning-capable AI agent that optimizes decision-making in dynamic environments using Monte Carlo Tree Search and Direct Preference Optimization (DPO). While not directly about stock prediction, their work illustrates the power of adaptive learning from environment interactions and feedback—principles that resonate with the idea of modeling ticker-specific reactions to market news .

A critical reflection on the agent paradigm comes from **Sayash Kapoor et al.**, whose work **"AI Agents That Matter"** critiques the overemphasis on accuracy without cost and contextual robustness. They advocate for jointly optimizing performance and efficiency, a notion relevant to stock market models that must balance predictive power with deployment cost and interpretability .

Bringing these diverse threads together, our proposed model bridges the domain-general strength of FinBERT with ticker-aware, sector-sensitive embeddings and end-to-end learnable sentiment fusion. Unlike previous models, our architecture dynamically adapts sentiment features through fine-tuning on Indian market data and explicitly models ticker identity. In doing so, we address

the core limitations of earlier works: the lack of localization, static sentiment modeling, and blindness to market-specific volatility.

This story of innovation—from static predictors to agentic, adaptable systems—reflects the ongoing transformation of financial forecasting from heuristic-driven tools to contextual, learning-based decision engines. Our contribution positions itself at the confluence of domain-aware NLP, deep time-series modeling, and financial context adaptation, offering a robust alternative to static, sentiment-agnostic models.

3 Data Understanding

3.1 Data Source

The datasets used in this project are sourced from two primary platforms:

• **yFinance API:**

Used to extract daily historical stock data, including open, high, low, close prices, and volumes for selected companies.

• **Kaggle:**

Provided the news headlines dataset with metadata such as publishing date, source, and sector tagging.

3.2 Data Description

The **Historical stock dataset** includes features such as:

• **Ticker:** Company symbol listed on BSE

• **Date:** Trading day

• **Open, High, Low, Close:** Daily stock prices

• **Volume:** Daily trading volume

• **Sector Name:** Categorization based on industry

Date	Open	Close	High	Low	Ticker	Sector
04-05-2022	3157.448168	3091.181396	3216.835908	3048.570645	GLAND.BO	Healthcare
05-05-2022	3104.98912	3135.524414	3158.635999	3091.923865	GLAND.BO	Healthcare
06-05-2022	3103.009423	3028.428223	3103.009423	2991.607872	GLAND.BO	Healthcare
09-05-2022	2991.162589	2997.200439	3011.552477	2940.683009	GLAND.BO	Healthcare
10-05-2022	3040.701939	2937.21875	3058.468724	2873.871825	GLAND.BO	Healthcare
11-05-2022	3009.96867	2864.369629	3009.96867	2819.878361	GLAND.BO	Healthcare
12-05-2022	2918.907585	2817.800049	2918.907585	2772.417818	GLAND.BO	Healthcare
13-05-2022	2820.917784	2964.833984	2992.152443	2820.917784	GLAND.BO	Healthcare
16-05-2022	2984.035996	3077.917969	3100.039998	2970.277691	GLAND.BO	Healthcare

Figure 1: Distribution of historical stock price data for selected companies.

The **news dataset** contains:

- **Date:** Headline publication date
- **Headline_category:** News Category
- **Headline:** News Headline

publish_date	headline_category	headline_text
20010102	unknown	Status quo will not be disturbed at Ayodhya; says Vajpayee
20010102	unknown	Fissures in Hurriyat over Pak visit
20010102	unknown	America's unwanted heading for India?
20010102	unknown	For bigwigs; it is destination Goa
20010102	unknown	Extra buses to clear tourist traffic
20010102	unknown	Dilute the power of transfers; says Riberio
20010102	unknown	Focus shifts to teaching of Hindi
20010102	unknown	IT will become compulsory in schools
20010102	unknown	Move to stop freedom fighters' pension flayed

Figure 2: Sample visualization of the financial news dataset used for embedding generation.

3.3 Data Preprocessing

Data preprocessing plays a pivotal role in preparing multi-modal data for deep learning models. For the stock dataset, we sorted records by Ticker and Date to maintain chronological sequence. A forward-shifted 'next_close' column was created to compute the target variable — the percentage change in stock price on the next trading day. The formula applied was:

```
stock_df["target_pct_change"] = ((stock_df["next_close"] - stock_df["Close"]) / stock_df["Close"]) * 100
```

For the news data, we first applied BART to classify and tag sector-specific headlines. Then, DistilBERT was used to transform each headline into a 768-dimensional CLS embed-

ding. These embeddings capture the sentiment and semantic meaning of the headlines in a high-dimensional space. To align news data with trading data, all embeddings corresponding to a Ticker-Date pair were averaged, resulting in one dense vector per trading day.

We categorized the news into sectors such as **Healthcare, Services, Financial Services, Consumer Discretionary, Information Technology etc.** , implementing Zero shot Classification using BART and the table is as follows:

publish_date	headline_category	headline_text	Predicted Sector
01-05-2022	india	Delhi: Over 1;500 Covid cases for 2nd day; positivity rate above 5%	Healthcare
01-05-2022	india	Cop kills self in Jammu's Arnia police station	Consumer Discretionary
01-05-2022	india	Devise ways to release 3.5 lakh undertrial prisoners: PM	Services
01-05-2022	india	Patiala clash: 3 senior cops moved; 3 accused held	Consumer Discretionary
01-05-2022	india	Hottest April on record in north; central India: IMD	Consumer Discretionary
01-05-2022	business.india-business	Rs 5;500 crore deposits of Xiaomi seized for 'forex violations'	Fast Moving Consumer Goods
01-05-2022	india	Pendency due to unclear laws; executive failure: CJI	Consumer Discretionary
01-05-2022	india	Gorakhnath attacker took IS oath in 2020?	Services
02-05-2022	business.india-business	Shriram City Union Fin hopes to grow AUM by 18%-20% in FY23	Financial Services

Figure 3: Distribution of financial news headlines across different industry sectors.

```python
from transformers import pipeline
import pandas as pd

# Load dataset
df = pd.read_csv("data_refined.csv")

# Load zero-shot classification pipeline
classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")

# Define candidate labels (sectors)
candidate_labels = [
    "Commodities",
    "Consumer Discretionary",
    "Fast Moving Consumer Goods",
    "Financial Services",
    "Healthcare",
    "Industrials",
    "Information Technology",
    "Services",
    "Telecommunication",
    "Utilities"
]
```

Figure 4: Zero-shot classification of headlines into industry sectors using BART.

```python
import pandas as pd

# Load your CSV file
df = pd.read_csv("news_with_sector.csv")

# Convert publish_date format
df['publish_date'] = pd.to_datetime(df['publish_date'], format='%d-%m-%Y').dt.strftime('%Y-%m-%d')

# Save it back to CSV
df.to_csv("news_with_sector1.csv", index=False,date_format='%Y-%m-%d')
```

Figure 5: Date formatting of news publish dates for consistency.

The final dataset merged these features using the Ticker and Date as keys. The closing

price was normalized using min-max scaling to ensure that all features were on comparable scales. The input features (CLS embeddings + normalized price) were reshaped into 3D tensors of shape (samples, 1, 769) to be LSTM-compatible.

We also removed dates on which the BSE was closed to ensure continuity in trading data and prevent the introduction of noise or temporal misalignment.In total, the dataset includes one year of consistent, high-quality data spanning from 2022 to 2023 across 53 actively traded companies, making it a robust foundation for building and evaluating our predictive framework.

4 Methodology

The methodology of this project is anchored in the interdisciplinary convergence of Natural Language Processing (NLP) and Time Series Forecasting, aiming to build a predictive framework that estimates stock price movements based on the semantic influence of financial news. The central hypothesis of our research is that financial headlines contain latent market sentiment cues which, when accurately decoded and fused with historical stock price trends, can offer predictive signals regarding short-term price fluctuations. This approach not only enhances the interpretability of stock behavior but also leverages the power of deep learning to model complex, non-linear relationships between qualitative news data and quantitative market behavior.

Our pipeline is built using a **multi-modal deep learning architecture** that integrates structured historical stock price data with unstructured textual data from financial news. This comprehensive framework is divided into several stages: **data collection, data preprocessing and temporal alignment, feature engineering, sentiment modeling with BERT, sector classification with BART, time-series modeling using LSTM, and real-time inference deployment.** Each phase contributes uniquely to the overarching goal of constructing a robust, real-world deployable prediction system.

The methodology is structured into several key stages:

1. Data Processing & Alignment

Synchronizing and merging stock price data with corresponding news headlines on a per-ticker,

per-day basis.

2. Sector Classification with BART

Employing a zero-shot classification model to accurately assign each headline to a relevant sector.

3. BERT Fine-Tuning for Impact Estimation

Training BERT model to learn the relationship between news headlines and subsequent stock price movements.

4. LSTM-Based Price Prediction

Integrating BERT-derived embeddings into a time-series LSTM model that predicts future price changes based on past prices and news context.

5. Performance Evaluation

Systematically assessing the model using standard regression metrics to validate its predictive power and generalizability.

This pipeline aims to simulate how market-moving news affects stock behavior, enabling a dynamic, data-driven forecasting system that reflects both quantitative patterns and qualitative insights.

4.1 Data Collection

The first step involved gathering and curating two key datasets that form the backbone of the model one structured and the other unstructured.

4.1.1 Historical Stock Data

This dataset encompasses daily stock price data for a range of publicly traded companies, spanning from **May 1, 2022, to May 1, 2023**. The data has been collected in a time-series format and includes multiple tickers representing various sectors within the Indian financial ecosystem.

The dataset comprises several key features that provide a comprehensive view of daily stock activity. Each entry includes the date, indicating when the stock was traded; the open price, representing the value at which the stock began trading on that particular day; and the close price, showing the final value by the end of the trading session. Additionally,

the ticker serves as a unique identifier for each company listed on the exchange. Another important attribute is the sector name, which classifies each company into categories such as Information Technology, Healthcare, Financial Services, and others, enabling sector-wise analysis and comparison.

This dataset forms the quantitative backbone of the model, representing market behavior over time. The core target variable used for prediction is the percentage change in stock price, calculated using the formula:

$$\text{Percentage Change} = \frac{\text{Close} - \text{Open}}{\text{Open}}$$

This metric reflects the intraday price movement of each stock and is treated as the output label in our supervised learning pipeline.

4.1.2 News Headline Dataset

Complementing the numerical stock data is a rich collection of **news headlines** published within the same time frame—i.e., from **May 1, 2022, to May 1, 2023.** These headlines represent real-world events, government announcements, earnings reports, mergers, policy changes, and other relevant occurrences that may influence investor sentiment and, by extension, stock prices.

The news dataset includes several important attributes that contribute to understanding its impact on market behavior. The published date specifies when each news headline was made public, aligning it with trading activity. The news headline itself is a concise, one to two-line summary that captures the core message or event. Additionally, the predicted sector indicates which segment of the economy is most likely to be affected by the news. This classification is generated using zero-shot learning with the BART (Bidirectional and Auto-Regressive Transformers) model, which assigns each headline to the most relevant sector based on predefined industry labels.

This dataset represents the **unstructured textual component** of the study and is critical

in capturing qualitative signals that may influence market behavior. Each headline, once classified, is paired with corresponding stocks from the same sector on the same date.

4.2 Data Preprocessing & Alignment

Raw data, while rich, often contains inconsistencies and gaps that need to be resolved before feeding it into deep learning models. To ensure model reliability and high-quality training samples, we applied a rigorous data preprocessing pipeline.

The Key steps are as following:

1. Date Standardization

All date fields in both datasets were standardized into a uniform format (YYYY-MM-DD) to facilitate proper merging and temporal alignment.

2. Handling Missing Values

In the stock price dataset, missing data points were forward-filled to maintain continuity without introducing artificial noise.

To fuse the insights from both datasets, a meticulous alignment and merging strategy was applied. For every news headline, the predicted sector was used to map the headline to all stocks in that sector. Then, the published date of the news was matched with the stock price data on the same date, allowing for the construction of a unified dataset where each sample includes:

• **A news headline** (unstructured textual input)

• **The corresponding sector** (used for filtering relevant stocks)

• The **stock's opening and closing prices** for the same day (used to compute the percentage change)

Figure 6: Data Preprocessing Flowchart

Date	headline_text	Sector	Open	Close	Ticker	price_change_pct
02-05-2022	Shriram City Union Fin hopes to grow AUM by 18%-20% in FY23	Financial Services	9.76	9.76	SSPNFIN.BO	0
02-05-2022	Shriram City Union Fin hopes to grow AUM by 18%-20% in FY23	Financial Services	11.02	11.47	KIDUJA.BO	0.040834827
02-05-2022	Shriram City Union Fin hopes to grow AUM by 18%-20% in FY23	Financial Services	3.75	3.54	OPTIFIN.BO	-0.05600001
02-05-2022	Global giants rediscover Chennai	Consumer Discretionary	3.04	3.04	INRADIA.BO	0
02-05-2022	Global giants rediscover Chennai	Consumer Discretionary	13.625	14.54167	CEENIK.BO	0.067278246
02-05-2022	Opposition bloc may not get BJD & YSR backing on Prez polls	Consumer Discretionary	3.04	3.04	INRADIA.BO	0
02-05-2022	Opposition bloc may not get BJD & YSR backing on Prez polls	Consumer Discretionary	13.625	14.54167	CEENIK.BO	0.067278246
02-05-2022	Was there when Babri razed; saw no Sena man: Fadnavis	Consumer Discretionary	3.04	3.04	INRADIA.BO	0
02-05-2022	Was there when Babri razed; saw no Sena man: Fadnavis	Consumer Discretionary	13.625	14.54167	CEENIK.BO	0.067278246

Figure 7: Merged Dataset by Ticker and Date

The result of these operations was a temporally aligned dataset where each ticker-day entry had both numerical stock price data and a corresponding news sentiment vector, ready for multi-modal modeling.

4.3 Feature Engineering

In this project, feature engineering plays a central role in bridging the gap between two inherently different data modalities: structured numerical stock data and unstructured textual news headlines. The objective is to construct a cohesive feature space that allows a deep learning model to capture both quantitative market patterns and qualitative market sentiment, thus enabling more accurate predictions of stock price changes.

4.3.1 Multi Modal Feature Set

The final model input is a multi-modal feature set, composed of:

1. Structured (Numerical) Features

The structured input primarily comprises daily stock trading metrics, which provide a historical quantitative foundation for modeling stock behavior. These features are critical for understanding price fluctuations based on past performance and general market activity.

The key components include:

- **Open Price:** The price at which a stock begins trading on a given day.
- **Close Price:** The final price of the stock at market close.
- **Percentage Change:** (Close price – Open price)/Open price

Although raw price values such as Open, Close, High, and Low are included as inputs, the primary target feature—percentage change—is already a normalized value. Hence, explicit normalization techniques like Min-Max Scaling or Z-score Normalization were not applied to the price features in the current version. However, ticker-wise variation is indirectly captured via the learnable ticker embeddings."

2. Unstructured (Textual) Features

To incorporate real-world market context and sentiment, the model integrates features derived from news headlines. These textual features are vital in modeling the causal or correlational influence of external news on intra-day price movements.

(a) News Headline Embeddings:

Each news headline is passed through a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model. Instead of using generic sentence embeddings, the BERT model has been fine-tuned on a custom task that maps headlines to percentage price changes, effectively learning a latent representation that reflects the effect of news on stocks. The final 768-dimensional [CLS] token embedding from BERT is extracted for each headline. These embeddings are semantically rich and encode complex syntactic, semantic, and contextual signals, allowing the model to understand nuances in news language (e.g., speculation vs confirmation, optimism vs caution).

(b) Sector Classification Scores:

To enhance the alignment between news and relevant stocks, we utilize a zero-shot classification approach using the BART model. Each news headline is evaluated against a predefined set of industry sectors (e.g., Information Technology, Healthcare, Financial Services), and a confidence distribution over sectors is generated.

These classification scores are included as additional features and serve to:

• Guide the model in determining which sector(s) the news is most likely to impact.

• Improve the precision of headline-to-ticker mapping, especially in cases where headlines may be ambiguous or cross-sectoral.

4.3.2 Combined Multi - Modal Input

By merging these components—numerical price data, ticker embeddings, contextual BERT embeddings, and sector classification scores—we create a comprehensive multi-modal input representation. This input is then reshaped and fed into a deep learning architecture (detailed in the next section), which learns to associate and weigh both structured market behavior and unstructured textual sentiment for predicting stock price changes.

This design ensures that the model:

- Understands market trends through time-series features.
- Learns ticker-specific behavior using ticker embeddings.
- Captures external signals and investor sentiment via headline embeddings.
- Aligns sector relevance using BART-based classification scores.

This thoughtful and nuanced feature engineering pipeline lays the groundwork for building a truly data-driven, real-world financial prediction system.

4.4 Sentiment Modeling with BERT

In financial markets, the sentiment embedded within news articles and headlines can have a significant impact on investor behavior and stock price volatility. To capture this latent signal, we designed and fine-tuned a custom BERT-based model to predict the numerical effect of each news headline on the corresponding stock's price movement, expressed as a **percentage change** from open to close on the same day. Unlike traditional sentiment classification (positive/negative/neutral), this model outputs a **continuous real-valued "impact score"**, thus serving as a proxy for how influential a piece of news is on a stock's performance.

4.4.1 Rationale for Using DistilBERT

While finance-specific models like FinBERT exist, they are heavily tailored to U.S. financial jargon and regulatory narratives. Our dataset includes Indian market news headlines spanning multiple sectors beyond finance (e.g., Healthcare, Services, Utilities). Therefore, we opted for **DistilBERT**, a distilled version of BERT, which balances semantic understanding with reduced computational overhead. DistilBERT retains 97% of BERT's language understanding while being 40% smaller and 60% faster.

Model Objective: Impact Regression, Not Sentiment Classification

Rather than simplifying sentiment into binary or categorical outcomes (e.g., positive/neutral/negative), this work reframes the problem as a regression task. The target variable is the percentage change in stock price, calculated as:

$$\text{Percentage Change} = \frac{\text{Close} - \text{Open}}{\text{Open}}$$

The model learns a continuous output that serves as a **proxy for the market impact of a given headline,** capturing the subtleties in language and how they may influence investor sentiment and price action.

Model Architecture Overview

The architecture consists of three major components:

- **Base Language Encoder:**

A pretrained distilbert-base-uncased model from Hugging Face, which captures the semantic structure of headlines. The [CLS] token output is treated as a compressed representation of the full headline.

- **Ticker Embedding Layer:**

Since different companies (tickers) may react differently to similar news, a learnable nn.Embedding layer was introduced. Each stock ticker is encoded as a trainable vector, allowing the model to differentiate behavior across stocks.

- **Regression Head:**

The final layers consist of fully connected feed-forward layers that process the concatenated BERT [CLS] output and ticker embedding. The output is a single continuous value predicting the price change percentage.

Activation function ReLU were experimented with to introduce non-linear learning capacity.

Fine-Tuning the Model

The DistilBERT model was fine-tuned end-to-end on a custom dataset comprising Indian

stock market news headlines and corresponding price movements. During this process, **all layers of the DistilBERT transformer were made trainable**, enabling the model to adjust its internal representations based on domain-specific language patterns and stock behavior.

This fine-tuning approach allowed the model to evolve beyond being a static semantic encoder. Instead, it became a **dynamic component**, capable of capturing how nuanced financial language impacts different sectors and ticker-specific price responses in the Indian market context.

By leveraging gradient-based optimization during training, the BERT embeddings were adapted to emphasize features most relevant to **market movement prediction**, rather than general linguistic understanding. This domain adaptation was particularly important, as the base DistilBERT model had been pretrained on generic English corpora, with no specialization in financial or India-specific news content.

Through this process, the model became more attuned to **market-moving phrases, sector-specific terminology, and variability in stock reactions**, improving its ability to generate meaningful impact signals from unstructured text.

Training Setup

- **Loss Function:** Mean Squared Error (MSE) was used to minimize the gap between predicted and actual percentage change values.
- **Optimizer:** Adam with a learning rate of 2e-5.
- **Epochs:** The model was trained over multiple epochs with early stopping enabled to prevent overfitting.
- **Batch Size:** 16
- **Validation Strategy:** An 80/20 train-test split ensured a fair evaluation on unseen data.
- **Evaluation Metrics:** MSE, RMSE, MAE, and R^2 score were tracked during training.

During training, we observed how the model's performance metrics—like Mean Absolute

Error and R^2—evolved over time, using these to guide model selection and early stopping.

Outcome of the Fine-Tuning Process

After training, the model was able to:

- Recognize impactful words and phrases in financial headlines.
- Learn ticker-specific sensitivity to news.
- Output a continuous signal that correlates with real-world price fluctuations.

This fine-tuned DistilBERT model, forms the foundation for real-time market news impact prediction, and is integrated as a primary input module for the LSTM-based architecture used in the final prediction pipeline.

4.5 LSTM-Based Time Series Model

Input Construction

The input to the LSTM model was a 3-dimensional array. The shape is given as (num_samples, sequence_length, feature_dim)

- **sequence_length:** The number of past days considered in each sample. We used a window of 7 trading days.
- **feature_dim:** Each daily record for a given ticker included:
  - A 768-dimensional [CLS] BERT embedding representing the semantic content of the news headlines for that ticker and day.
  - A single scalar value indicating the same-day price change percentage (used as a feature in the sequence).
  - This resulted in a feature vector of dimension 769 per day.
- **LSTM Layers:** Two stacked LSTM layers with a hidden size of 64 units. The final hidden state of the sequence was passed to the next layer.

• **Dense Regressor:** A fully connected linear layer that maps the LSTM output to a single scalar value representing the predicted price change percentage.

• **Loss Function:** Mean Squared Error (MSE) was used to measure the difference between predicted and actual values.

• **Optimization:** Adam optimizer with a learning rate of 1e-3.

• **Regularization:** Dropout and layer normalization were used to mitigate overfitting.

Training and Evaluation

The model was trained on 80% of the total dataset, while 20% was reserved for . No shuffling was applied to preserve the sequential nature of the data.

We evaluated the model on standard regression metrics:

• Root Mean Squared Error (RMSE)

• Mean Absolute Error (MAE)

• R-squared (R^2)

5 Results and Discussion

5.1 Model Performance and Evaluation

The proposed BERT-LSTM model was evaluated on a real-world dataset combining news headlines, historical stock prices, and ticker-level information. The dataset was split in an 80:20 ratio for training and validation, respectively, with no shuffling to maintain temporal consistency—a critical consideration for time series modeling.

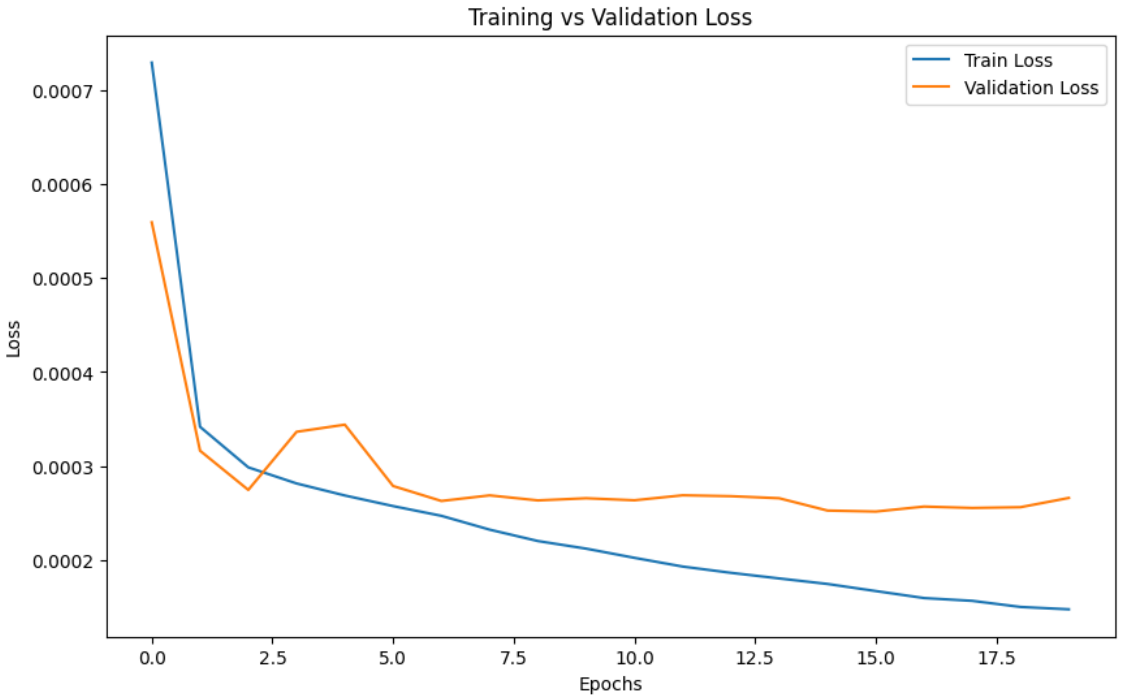

Figure 8: Training vs validation loss across epochs.

The model exhibited stable and progressive convergence, as evidenced by a consistent decline in training and validation loss over time. Initially, the model rapidly learned the basic structure of the data, reflected in a sharp improvement across all performance metrics. As training progressed, the loss curves plateaued, indicating that the model had entered a regime of stable generalization, capturing complex relationships between input features and the target variable.

Notably, the validation performance stabilized without significant fluctuations, suggesting that the model was not overfitting despite the inclusion of high-dimensional BERT embeddings. The use of layer normalization, dropout, and a moderately deep LSTM architecture helped the model retain generalizability across unseen data.

Performance Evaluation

On the validation set, the model achieved the following final results:

- Root Mean Squared Error (RMSE): 0.0162
- Mean Absolute Error (MAE): 0.0078

- R-squared (R^2): 0.7192

These metrics indicate a strong predictive capability, particularly in a highly volatile domain such as stock price forecasting. The low RMSE and MAE values highlight the model's precision in estimating price change percentages, while the R^2 score confirms that the model was able to explain a substantial proportion (72%) of the variance in the target variable.

Interpretation

The results confirm that integrating semantic representations of news (via BERT), temporal price history (via LSTM), and ticker-specific embeddings enables the model to make nuanced predictions about short-term stock movements. The ability to capture subtle shifts in market sentiment and contextual relevance to specific stocks offers a significant improvement over traditional price-only models or naive sentiment tagging.

Overall, the training and evaluation outcomes support the effectiveness of the proposed architecture for time-sensitive, context-aware stock prediction tasks.

- Model Performance

  Compare different models based on evaluation metrics.

- Key Insights

  Interpret model predictions and real-world implications.

5.2 Sector Prediction from Financial News Headlines

To evaluate the system's ability to classify the correct stock sector from financial news headlines, sample predictions were tested using real-world-like headlines.

Example 1:

News Headline: "RBI cuts repo rate by 0.25%, trims GDP growth forecast Repo Rate"Predicted

Sector: Financial Services

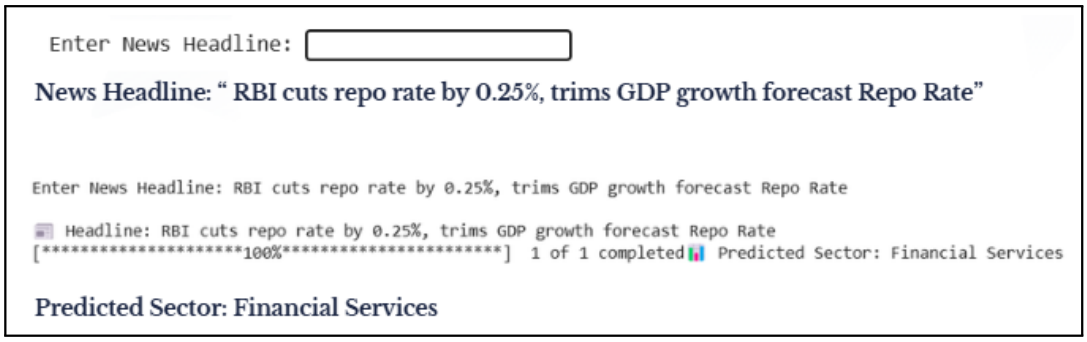

Figure 9: Financial Services Headline Prediction

Example 2:

News Headline:"Trump to Impose 104% Tariffs on All Chinese Imports, Escalating Trade Tensions over Beijing's Retaliatory Tariff Measures"

Predicted Sector: Industrials

Figure 10: Industrials Headline Prediction

These examples demonstrate the model's capability to associate macroeconomic cues and geopolitical signals with the correct industrial classification.

5.3 Price Percentage Change Prediction Results

For each correctly identified sector, the model further predicted the stock price percentage change of relevant companies (based on tickers within the sector).

Ticker	Predicted % Change	Actual % Change	Accuracy Indicator
KHANDSE.BO	-0.0096	-0.0426	Close match
RAJPUTANA.BO	-0.0392	-0.0489	Close match
KEYFINSERV.BO	-0.0222	0.0613	Opposite trend
Ticker 3	-0.004	0	Neutral outcome

Figure 11: Financial Sector Predictions

Ticker	Predicted % Change	Actual % Change	Accuracy Indicator
MADHUCON.BO	0.0005	0.0732	Close Match
PARAS.BO	-0.0094	0.014	Opposite trend

Figure 12: Industrial Sector Predictions

Interpretation

• The sector classification model shows high contextual awareness, accurately mapping diverse news headlines to relevant sectors.

• The price prediction model demonstrates partial accuracy, with several predictions falling close to actual changes. However, volatility or macro noise may have influenced discrepancies, particularly in the Industrials sector.

6 Conclusion and Future Work

6.1 Summary of Findings

This capstone project presents a multi-modal deep learning framework designed to predict short-term stock price fluctuations by integrating general news sentiment with historical stock data. Rooted in the hypothesis that news headlines—even those not explicitly financial—can influence market behavior, this study explores the fusion of unstructured textual signals and structured time-series data to model stock movements more holistically.

Using DistilBERT, we generated sentiment embeddings from daily news headlines and aligned them with historical stock data sourced from the Bombay Stock Exchange (BSE) for 53 actively traded companies across ten industry sectors. To improve sector-specific relevance, BART-based zero-shot classification was used to map each headline to a corresponding sector. These sentiment features, combined with normalized stock prices, were fed into an LSTM-based time series model to predict the next day's percentage change in stock prices.

The model was trained and validated on data spanning 2022 to 2023. Evaluation metrics on the validation set—such as a Root Mean Squared Error (RMSE) of 0.0162, Mean Absolute Error (MAE) of 0.0078, and an R^2 score of 0.7192—demonstrate the model's strong predictive perfor-

mance in capturing both sentiment-driven and temporal dynamics.

This project underscores the feasibility and value of using general news signals for real-time stock prediction. By bridging qualitative public sentiment and quantitative market data, our approach offers a scalable and insightful tool for stock trend forecasting. Future enhancements may include expanding the prediction window, refining sector-specific modeling, and deploying the system for real-world market applications.

6.2 Limitations

While the proposed BERT-LSTM-based multi-modal architecture demonstrated promising results in predicting short-term stock price movements using news and historical data, several inherent limitations exist in the current approach that warrant discussion. These limitations span modelling choices, data representation, labelling methodology, and practical deployment concerns.

- **Absence of Lagged News Impact Modelling**

  A fundamental assumption in this work is the immediate effect of news on stock prices, wherein sentiment is inferred using the same-day price change and predictions are also made for the same day. This disregards the potential lagged effect that news can exert on market behaviour. Investor reactions may unfold over multiple trading sessions due to delayed interpretation, institutional decision-making cycles, or liquidity constraints. The model, by not accounting for such temporal spillovers, might miss capturing the extended influence of impactful news or overestimate short-term reactivity.

- **Limited Modelling of Sectoral and Macroeconomic Context**

  Although sector classification was incorporated using a zero-shot BART-based classifier, the model does not explicitly account for **sector-level interdependencies or macroeconomic indicators**. Stock movements are often not isolated but influenced by broader industry trends, policy changes, or global events. The model's ticker-level design may therefore miss systemic risks or opportunities arising from sector-wide dynamics, resulting in underfitting for correlated movement patterns among companies within the same industry group.

- **Sentiment Labelling via Price Change Proxy**

The sentiment model was supervised using a **regression target derived from same-day price percentage change**, assuming that positive price movement reflects positive sentiment and vice versa. This indirect labelling introduces noise, as price movements are often influenced by **non-news factors** such as technical trading signals, liquidity pressures, or broader market sentiment. Moreover, some impactful news may not be priced in immediately or at all. Thus, the sentiment signal learned may conflate genuine semantic tone with market artifacts, limiting the model's ability to generalize sentiment patterns across contexts.

- **Dependency on Data Quality and Coverage**

  The reliability and representativeness of the input data play a crucial role in the performance of any data-driven model, particularly in the financial domain where market sensitivity to information is high. A significant limitation of the present work lies in the quality and coverage of the news data used for training.

  The news dataset employed in this study was sourced from a publicly available Kaggle repository, which, while convenient for experimentation, lacks the editorial rigor, verification protocols, and relevance filtering typical of professional news services. As a result, the dataset contained a heterogeneous mix of high-impact and low-impact articles, including general sector commentary, redundant items, and even non-relevant content, much of which is unlikely to influence market behaviour. This introduced substantial noise into the feature space, diluting the semantic signals essential for accurate sentiment extraction and stock price movement prediction.

  In contrast, financial intelligence providers such as **Bloomberg, Reuters, or S&P Global Market Intelligence** offer curated and timestamped news data that aligns more closely with actual investor behaviour and trading reactions. These sources often include metadata such as relevance scores, sentiment tags, event type annotations, and links to financial instruments, all of which are valuable in building robust, real-time market prediction systems.

6.3 Future Enhancements

While the current model demonstrates promising capabilities in integrating textual and quantitative data for short-term stock price movement prediction, there are several directions for future

enhancement. First and foremost, using high-quality, real-time financial news feeds from authoritative sources such as Bloomberg, Reuters, or S&P Global Market Intelligence would substantially improve input data fidelity. This would help ensure higher relevance, timeliness, and reliability of news content, thereby boosting the accuracy of sentiment extraction and its correlation with market movements.

Secondly, the sentiment labelling methodology can be improved. Currently, sentiment is inferred indirectly from same-day price movement, which overlooks lag and causality. Future iterations could incorporate market reaction delays and use annotated sentiment labels from Stock market information sources or other reliable sources, enabling the training of more semantically accurate models. Furthermore, expanding the sentiment classification from a simple regression-based encoding to multi-label or aspect-based sentiment analysis could capture richer dimensions of news influence.

Another important direction is the incorporation of time-aware mechanisms such as transformers with positional encoding or temporal attention layers, which can better model long-range dependencies and asynchronous effects between news and market behavior. Additionally, leveraging multi-task learning by simultaneously predicting volatility, trading volume may yield richer representations and improved generalization.

Future work can incorporate macroeconomic indicators, sector performance trends, and global market signals to provide a more holistic understanding of stock behavior. Finally, scaling the architecture to a real-time, deployable pipeline that can ingest live news, generate sentiment embeddings, and update predictions on-the-fly remains a compelling goal for practical application in algorithmic trading or investor decision-support systems.

7 References

• *Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, Rafael Rafailov. (2024). Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents.*

• *Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, Arvind Narayanan. (2024). AI Agents That Matter.*

• *Rahul Jain, Rakesh Vanzara. (2023). Emerging Trends in AI-Based Stock Market Prediction: A Comprehensive and Systematic Review.*

• *Madhusmita Khuntia, Deepa Gupta. (2023). Indian News Headlines Classification using Word Embedding Techniques and LSTM Model.*

• *Narayana Darapaneni, Anwesh Reddy Paduri, Himank Sharma, Milind Manjrekar, Nutan Hindlekar, Pranali Bhagat, Usha Aiyer, Yogesh Agarwal. (2021). Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets.*

• *(N.D.). FinBERT-LSTM Model for Stock Market Prediction Using Sentiment Analysis and Historical Data.*