Text Representation For Direction Prediction Of Share Market

BY :

Ashutosh Anand 202318035

Kunal Anand 202318057

Anjali Singh 202318050

INTRODUCTION

Financial news articles report on critical economic events, such as policy declarations, market trends, and commodity trading, which significantly influence stock market movements. This project investigates the relationship between financial news articles and stock price movements, focusing on predicting the direction of close price movement. We explore the effectiveness of using financial news articles alone, historical prices alone, and a combination of both for prediction.

To address this, we propose a parallel CNN model that integrates the text from financial news articles with historical prices and technical indicators, capturing both explicit and implicit relationships within the data. By focusing on news published during market hours, the study highlights the short-term impact of financial events on stock market direction, validated on key Indian indices: NIFTY 50, NIFTY Next 50, and NIFTY Bank (2010–2021).

Dataset

The project uses historical price data of NIFTY 50, NIFTY Next 50, and NIFTY Bank, along with financial news articles from 2010 to 2021, comprising 2978 trading days per index.

	Year	News article count (Pre-filtering)	News article count (Post-filtering)	No. of trading days
0	2010	16696	7844	252
1	2011	21942	10322	247
2	2012	24272	11409	251
3	2013	31333	14246	250
4	2014	41824	16568	244
5	2015	47955	17878	248
6	2016	37599	16524	247
7	2017	41159	22663	248
8	2018	34330	20748	246
9	2019	37641	24587	245
10	2020	50424	31881	252
11	2021	45221	21070	248
12	Total	430396	215740	2978

For the news dataset, archived articles from Economic Times between 2010 and 2021 were used. Each article includes a headline, publish date, and article body. Headlines published during active market hours (9 AM to 4 PM IST) were selected for the experiments.

A year-wise breakdown of the dataset shows the number of articles before and after filtering, along with the distribution of trading days per year. This dataset serves as the basis for analyzing the short-term impact of financial news on stock price movement prediction.

_id	Date	Title
0c527fa7e	2010-01-0	Companies adopt wierd names to catch consumer attention
b4ef0747e	2010-01-0	3 Idiots set to be Bollywoodâ€™s biggest grosser
495032c67	2010-01-0	Bharti bags Bangladeshâ€™s Warid for bargain price of Rs 45 la
ae0f4e9d7	2010-01-0	Corporate air travel takes off once again
1a4fbde9c	2010-01-0	Global mediclaim policy awaits jet-setting executives
3747fdff1(	2010-01-0	Shree Cement acquires 90 acres for new projects
65b9cff2cl	2010-01-0	Query Corner: Insurance
0ee85fd4c	2010-01-0	DoT plans country-wide MNP launch by March 31
b400300bf	2010-01-0	Cooking oil price may rise on low production
67740de5(	2010-01-0	Maharashtra, Gujarat in race for UP sugar processing deal
d27befec5	2010-01-0	Pranab Mukherjee plans more cash flow for farm growth
f20087ab8	2010-01-0	Fannie, Freddie proving too big to shrink
8ca8d37be	2010-01-0	I-T dept hits upon bogus investment losses
06d26749(	2010-01-0	â€˜Indian cos need to intensify sales and marketing effortsâ€™
5aee9704(	2010-01-0	India may ask US to ease trade curbs in fresh round of talks
dc6f51531	2010-01-0	General insurers to set up Motor Insurance Bureau
455bd7a7(	2010-01-0	PMO steps in after BSNL puts bid on hold
c93ffb724(	2010-01-0	Delta, Northwest can work as single carrier
5aa75138(	2010-01-0	BSNL may have to shelve $1-bn IT outsourcing deal
edb9a0af1	2010-01-0	Itâ€™s full steam ahead for Tinplate
7b758028(	2010-01-0	Q2 current account gap stays flat at $12.6 billion
b1686d25(	2010-01-0	Re climbs 5%, call rates hit by liquidity
e1f72d032	2010-01-0	Oil & gas regulator slaps notice on Centre, PSUs on fuel pricing
11a9511f1	2010-01-0	Anil Kumar may plead guilty in the Galleon hedge fund case
5cd7e387(	2010-01-0	Cutting deals is an art this globe trotter has perfected
cd37f1b61	2010-01-0	Mega road projects may be open only to big boys
c300e1e67	2010-01-0	Flush with funds, going gets easier
5fecc49e3	2010-01-0	REC eyes 26% in coal mines, power plants

Dataset

The project also uses historical price data of NIFTY 50, NIFTY Next 50, and NIFTY Bank .
We also calculate the following technical indicators: Average Directional Index (ADX), Moving Average Convergence Divergence (MACD), Momentum (MOM), Average True Range (ATR), Relative Strength Index (RSI), Stochastic Oscillator (STOCH), Bollinger Bands (BBANDS), Exponential Moving Average (EMA), Simple Moving Average (SMA) for all the indices.
The class labels for the direction of the close price movement are derived as follows:

- If the close price on day t (ct) is greater than or equal to the close price on day t-1(ct-1), the label t=1
- Otherwise, label t=0.

Date	Open	High	Low	Close
2-Jan-10	4882.05	4918.8	4827.15	4899.7
4-Jan-10	5249.2	5298.6	5249.2	5290.5
6-Jan-10	5086.25	5086.95	4961.05	4970.2
7-Jan-10	5312.05	5312.55	5232.1	5251.4
9-Jan-10	5403.05	5478.6	5403.05	5471.85
10-Jan-10	6030.3	6153.1	6030.3	6143.4
11-Jan-10	6092.3	6132.4	6084.75	6117.55
12-Jan-10	5871	5971	5865.55	5960.9
13-Jan-10	5212.6	5239.2	5169.55	5233.95
14-Jan-10	5234.5	5272.85	5232.5	5259.9
15-Jan-10	5259.9	5279.85	5242.45	5252.2
20-Dec-21	16824.25	16840.1	16410.2	16614.2
21-Dec-21	16773.15	16936.4	16688.25	16770.85
22-Dec-21	16865.55	16971	16819.5	16955.45
23-Dec-21	17066.8	17118.65	17015.55	17072.6
24-Dec-21	17149.5	17155.6	16909.6	17003.75
27-Dec-21	16937.75	17112.05	16833.2	17086.25
28-Dec-21	17177.6	17250.25	17161.15	17233.25
29-Dec-21	17220.1	17285.95	17176.65	17213.6
30-Dec-21	17201.45	17264.05	17146.35	17203.95
31-Dec-21	17244.5	17400.8	17238.5	17354.05

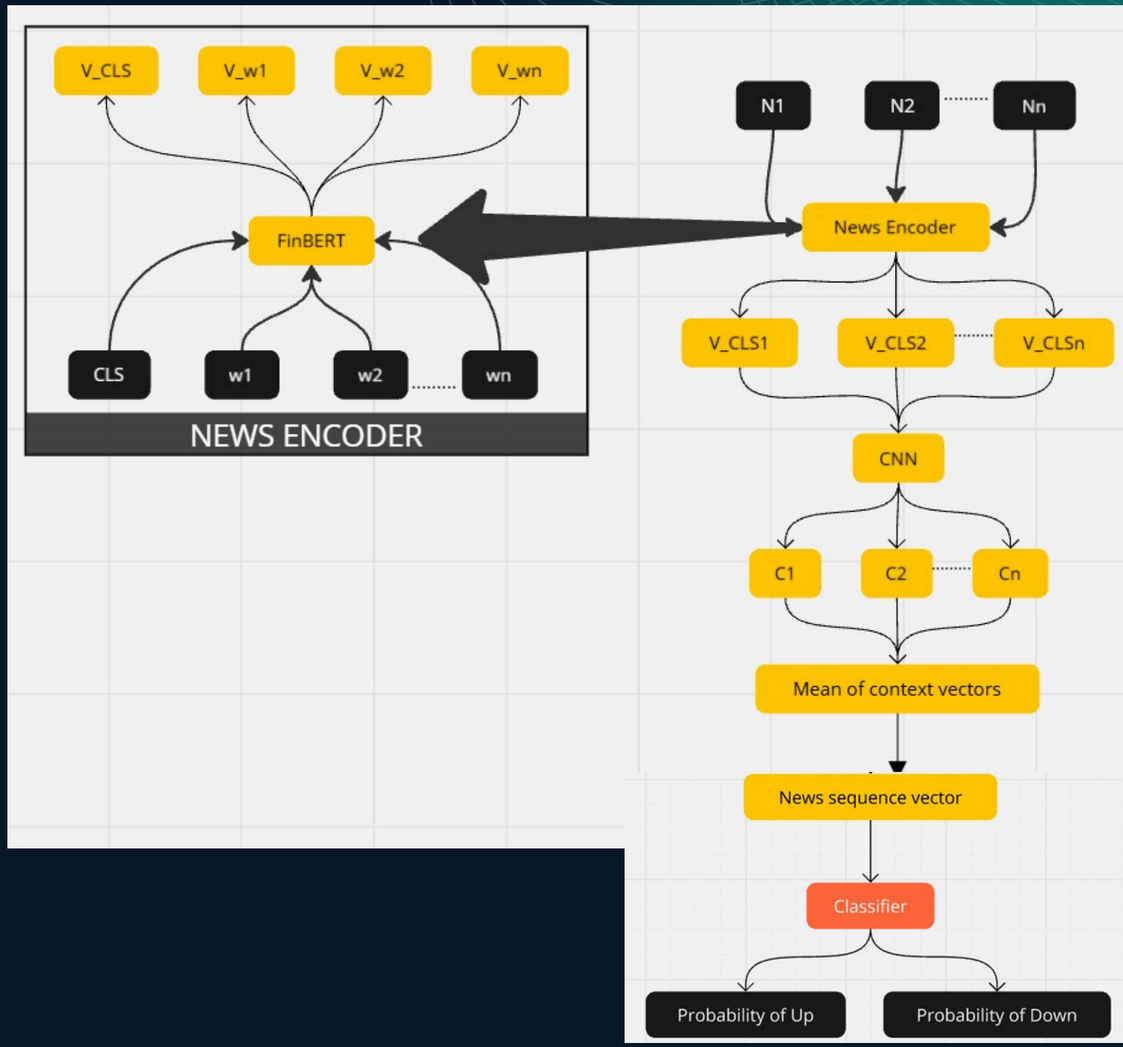
	Date	Open	High	Low	Close	ADX	MACD	MACD_Sig	MACD_Hist	MOM	...	RSI	SlowD	SlowK	WILLR	Upper_Band	Middle_Band	Lower_Band	SMA	EMA	Label
33	22-Feb-2010	10081.00	10139.85	9933.95	9948.50	13.390135	-312.553386	-204.408194	-108.145192	-1291.45	...	44.770598	9.738036	8.468672	-99.570986	10263.068918	10081.18	9899.291082	10941.971667	10803.245071	0
34	23-Feb-2010	9941.15	10002.30	9903.15	9928.25	13.336128	-334.535186	-230.433592	-104.101594	-1770.50	...	44.613742	10.398947	10.210952	-99.266575	10261.724736	10046.13	9830.535264	10873.473333	10746.793776	0
35	24-Feb-2010	9897.50	9951.95	9871.20	9936.60	13.324550	-347.278930	-253.802660	-93.476271	-2102.20	...	44.699778	9.825868	9.987617	-98.106680	10142.379227	9991.52	9840.660773	10776.481667	10694.523209	1
36	25-Feb-2010	9962.45	9981.75	9904.35	9918.35	13.263921	-354.761591	-273.994446	-80.767145	-3389.55	...	44.536938	14.172231	11.465682	-98.635015	10017.680859	9949.29	9880.899141	10666.831667	10644.447519	0
37	26-Feb-2010	9931.85	10181.00	9919.55	10099.95	12.865873	-342.094598	-287.614476	-54.480122	-2583.95	...	46.620783	36.416736	20.138278	-93.377723	10101.409985	9966.33	9831.250015	10582.253333	10609.318646	1

5 rows × 21 columns

Text Only Approach

Relies solely on news headlines published when the stock market is open. Assumes that market-relevant information contained in these headlines can provide enough insight to predict stock movements.

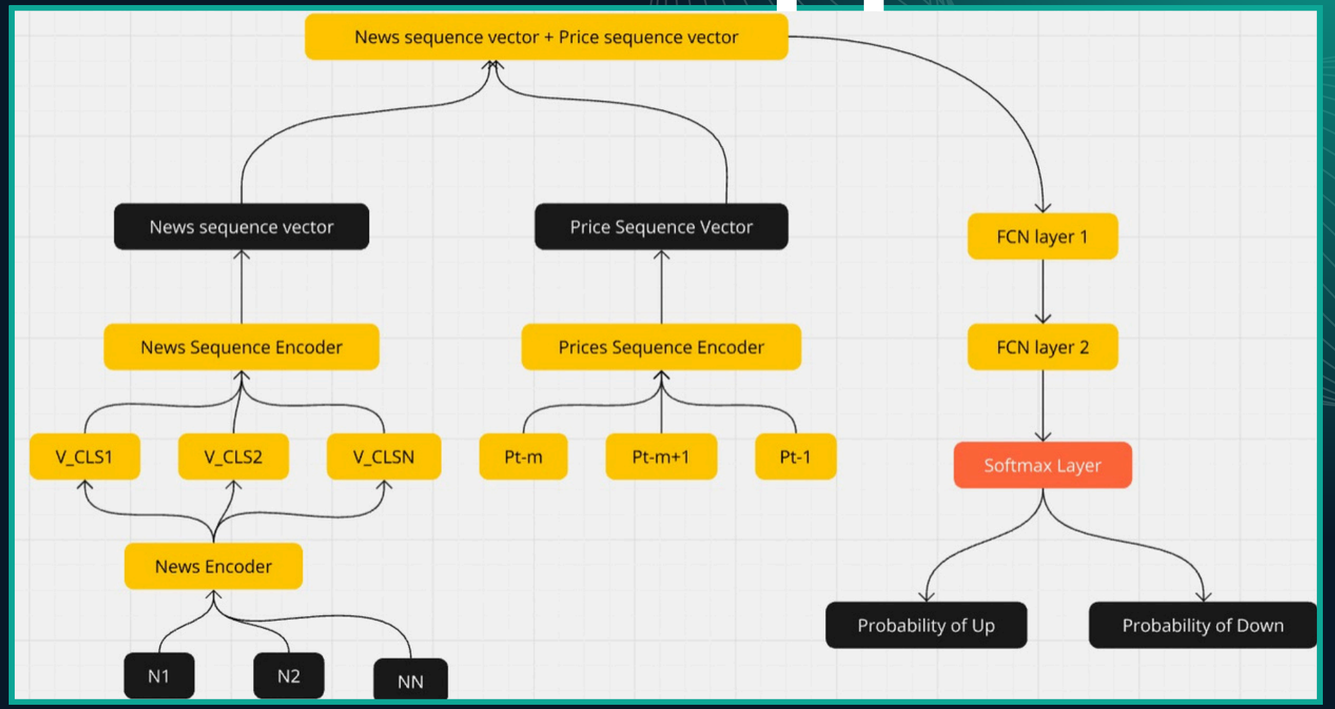
1. News Encoder:
   - Converts news headlines into meaningful vectors.
   - Steps:
     - Each headline is split into a sequence of words.
     - Words are encoded using FinBERT (a language model pre-trained on financial text).
     - FinBERT outputs:
       - The CLS vector (a summary representation of the headline).
       - Word-specific vectors.
     - The CLS vector is used as the headline's representation.
2. News Encoder:
   - Aggregates multiple headline vectors into a news sequence vector for the day.
   - Steps:
     - Passes the sequence of headline vectors through a CNN layer.
     - CNN captures local relationships between consecutive headlines.
     - Outputs context vectors that summarize daily news trends.
     - These vectors are averaged to create a single news sequence vector for the day.
3. Classifier:
   - Uses the news sequence vector to predict price movement direction.
   - Compares the performance of seven classifiers:
     - Random Forest, SVM, Perceptron, LSTM, CNN, etc.
   - Outputs:
     - The probability of the direction being up or down.

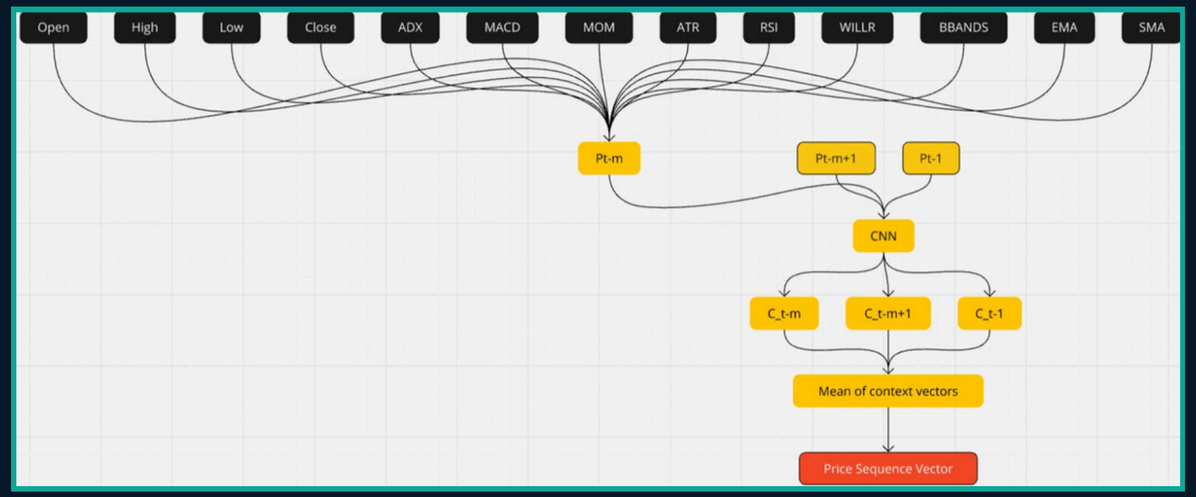


Combining Text and Price Approach

Integrate news headlines, historical prices, and technical indicators into a single model for better predictions.

1. News Encoder (Same as Text-Only Approach):
   - Encodes individual headlines into CLS vectors using FinBERT.
   - Aggregates multiple headlines into a news sequence vector.
2. Prices Sequence Encoder:
   - Encodes historical price data and technical indicators into vectors.
   - Steps:
     - Input: Historical prices and indicators (e.g., ADX, RSI, MACD).
     - Pass this data through a CNN layer.
     - CNN captures local relationships between consecutive days.
     - Outputs context vectors summarizing trends in price data.
     - These vectors are averaged to form a single price sequence vector.
3. Prediction Layer:
   - Combines the news sequence vector and price sequence vector.
   - Steps:
     - Concatenate the two vectors.
     - Pass through two fully connected layers.
     - Apply a softmax function to output probabilities for price direction (up or down).





Experimentation & Evaluation

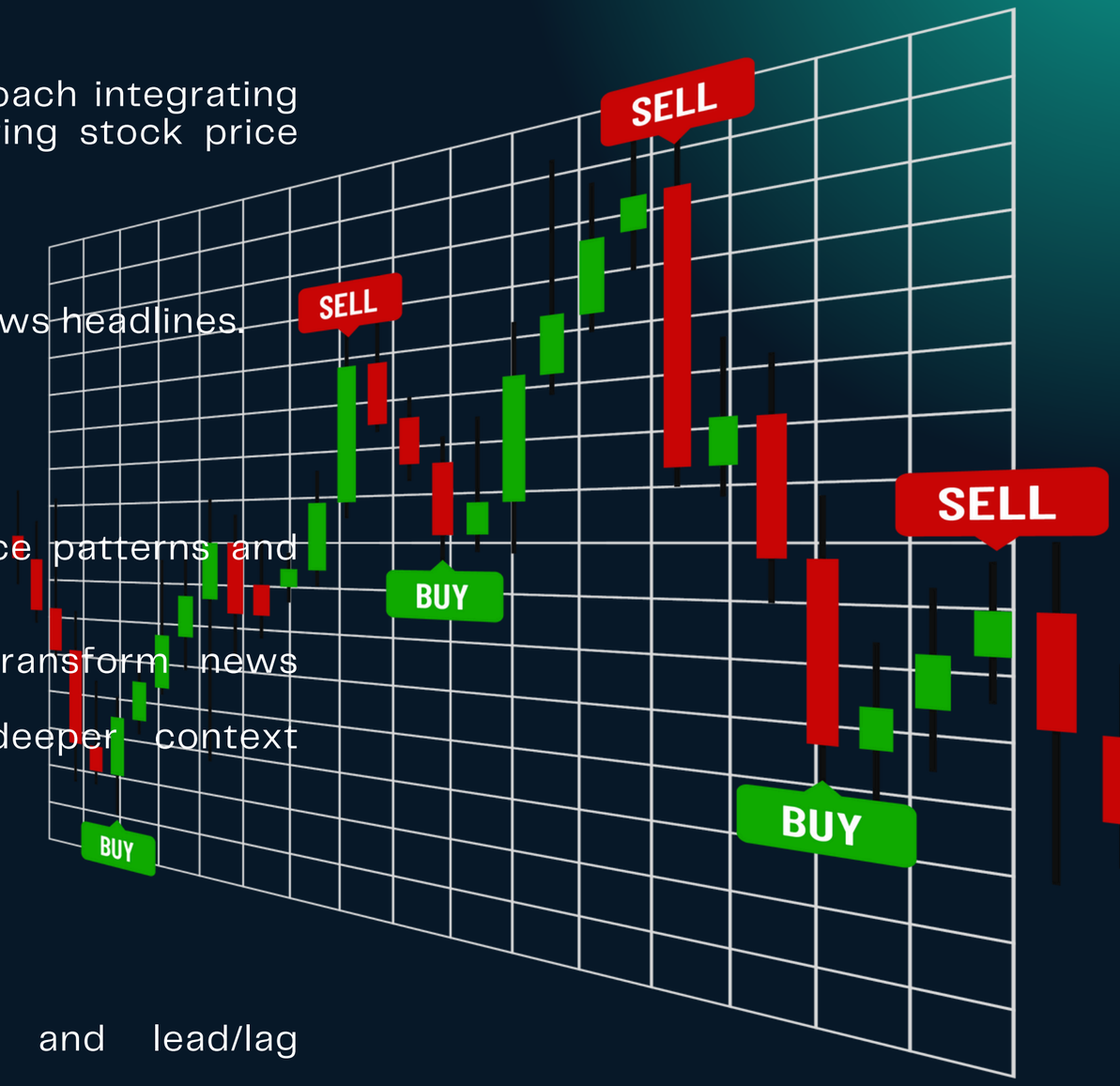
This section highlights the baseline models, a combined approach integrating text and price data and evaluation metrics used for predicting stock price movement .

Baseline Models

- Models predict stock price movement using price data or news headlines.
- Serve as benchmarks for evaluating advanced approaches.
- Price-Based Models:
  - SVM : Linear kernel , 10-day price vectors.
  - Random Forest : 100 decision trees.
  - Perceptron : Trained for 100 iterations.
  - LSTM , CNN , Transformer : Capture sequential price patterns and context.
- Text-Based Models:
  - SVM, Random Forest, Perceptron: Use TF-IDF to transform news headlines into features.
  - LSTM, CNN, Transformer: Use BERT/FinBERT for deeper context encoding.

Evaluation

- Evaluation Metrics :
  a. Accuracy : Measures correct predictions.
  b. ROC-AUC : Assesses discriminative power of models.
  c. Normalized Cross-Correlation: Captures alignment and lead/lag relationships between predictions and actual data.



Results : Price only approach

Performance Metrics of Various Classifiers for NIFTY 50 Price Prediction

	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	SVM	0.4936	0.5003	0.2459	6
1	Random Forest	0.4347	0.5098	0.1549	11
2	Perceptron	0.4181	0.4984	0.0851	-235
3	LSTM	0.4988	0.5013	0.2006	32
4	CNN	0.5043	0.5019	0.1565	12
5	Transformer Encoder	0.5118	0.5000	0.0148	-231

Performance Metrics of Various Classifiers for NIFTY NEXT 50 Price Only Prediction

	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	SVM	0.5336	0.5290	0.2717	4
1	Random Forest	0.4247	0.4618	0.1474	-44
2	Perceptron	0.4181	0.5004	0.0051	-237
3	LSTM	0.5588	0.4885	0.0148	-88
4	CNN	0.5043	0.4915	0.1638	11
5	Transformer Encoder	0.5498	0.4835	0.0290	-233

Performance Metrics of Various Classifiers for NIFTY BANK Price Only Prediction

	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	SVM	0.4864	0.5189	0.1848	18
1	Random Forest	0.4715	0.5200	0.1602	-34
2	Perceptron	0.4581	0.4980	0.0191	-235
3	LSTM	0.5388	0.4905	0.1559	-67
4	CNN	0.4543	0.5000	0.1517	-152
5	Transformer Encoder	0.5418	0.5040	0.0038	-241

This approach uses historical price data and technical indicators to predict stock price direction.
Transformer Encoder showed the highest accuracy across NIFTY 50 and NIFTY Bank.
However, its ROC-AUC score was 0.5 for all indices, indicating it randomly assigned labels.
Other Models' Performance:
Random Forest achieved the highest ROC-AUC for NIFTY 50 , NIFTY BANK .
SVM had the best ROC-AUC for NIFTY Next 50 , again marginally above random.
NCC scores showed weak correlation between predicted and actual series, with observable lag or lead in predictions .Thus we can infer that the prediction models cannot discover any pattern from the price representation approach.

```
Evaluation of model for year 2017
Accuracy 0.37815126050420167
Lag: -237 and cross correlation: nan
ROC AUC score: 0.5
Evaluation of model for year 2018
Accuracy 0.4872881355932203
Lag: -235 and cross correlation: nan
ROC AUC score: 0.5
Evaluation of model for year 2019
Accuracy 0.4829059829059829
Lag: -233 and cross correlation: nan
ROC AUC score: 0.5
Evaluation of model for year 2020
Accuracy 0.4066390041493776
Lag: -91 and cross correlation: 0.15945663030824592
ROC AUC score: 0.5029199711607786
Evaluation of model for year 2021
Accuracy 0.40756302521008403
Lag: -237 and cross correlation: nan
ROC AUC score: 0.5
```

Results : Text only approach

Performance Metrics of Various Classifiers for NIFTY 50 Text Only Prediction

	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	TF-IDF	SVM	0.6551	0.6362	0.3913	-22
1		Random Forest	0.6310	0.6143	0.3564	-4
2		Perceptron	0.6373	0.6207	0.2678	10
3	BERT	LSTM	0.7011	0.6822	0.4446	0
4		CNN	0.7174	0.6997	0.4239	0
5		Transformer Encoder	0.7098	0.6763	0.3846	0
6	FinBERT	LSTM	0.7091	0.6993	0.3501	0
7		CNN	0.7212	0.7005	0.4360	0
8		Transformer Encoder	0.6890	0.6596	0.3970	0

Performance Metrics of Various Classifiers for NIFTY Next 50 Text Only Prediction

	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	TF-IDF	SVM	0.6312	0.6198	0.2416	-5
1		Random Forest	0.6084	0.5728	0.1750	-20
2		Perceptron	0.6122	0.6113	0.3051	-4
3	BERT	LSTM	0.6664	0.6355	0.3339	0
4		CNN	0.6570	0.6410	0.2911	0
5		Transformer Encoder	0.6753	0.6777	0.3504	0
6	FinBERT	LSTM	0.6863	0.6667	0.4287	0
7		CNN	0.6402	0.6256	0.3318	0
8		Transformer Encoder	0.6665	0.6560	0.3222	0

Performance Metrics of Various Classifiers for NIFTY Bank Text Only Prediction

	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	TF-IDF	SVM	0.62080	0.6217	0.1978	2
1		Random Forest	0.59970	0.5965	0.2721	4
2		Perceptron	0.61337	0.5881	0.2889	-10
3	BERT	LSTM	0.68410	0.6549	0.4035	0
4		CNN	0.68530	0.6755	0.3442	0
5		Transformer Encoder	0.67860	0.6502	0.2986	0
6	FinBERT	LSTM	0.69960	0.6969	0.3892	0
7		CNN	0.69600	0.6662	0.3514	0
8		Transformer Encoder	0.66630	0.6353	0.3085	0

SVM with TF-IDF : Highest accuracy and ROC-AUC among traditional machine learning models.
Encoder-Based Models : Improved accuracy and ROC-AUC compared to non-encoder models.
FinBERT with CNN: Best for NIFTY 50 and NIFTY Bank.
FinBERT with LSTM: Best for NIFTY Next 50, with slightly lower ROC-AUC but higher accuracy than BERT with Transformer .

NCC Performance :
SVM with TF-IDF : Initial NCC scores (2017) showed lag, with weak correlation to actual values . However this does not hold over time indicating that there may exist some implicit relationship between the news articles that SVM cannot model.
Encoder-Based Models:
Showed stronger correlations with no observable lag or lead.
FinBERT-based models consistently achieved the highest NCC scores .

Key Observations :
Textual features improved accuracy, ROC-AUC, and NCC compared to price-only approaches.
Encoder-based models successfully modeled implicit relationships in news headlines, outperforming traditional methods.

```
[ ] Loading Train and Test Data ...
100%|██████████| 1738/1738 [08:02<00:00,  3.61it/s]
100%|██████████| 248/248 [01:10<00:00,  3.53it/s]
Prediction Model SVM
Evaluation of year: 2017
Accuracy 0.5685483870967742
Lag: 1 and cross correlation: 0.18255195698641116
ROC-AUC score 0.5301075268817205
Loading Train and Test Data ...
100%|██████████| 1986/1986 [09:12<00:00,  3.60it/s]
100%|██████████| 246/246 [01:10<00:00,  3.51it/s]
Prediction Model SVM
Evaluation of year: 2018
Accuracy 0.5203252032520326
Lag: 66 and cross correlation: 0.1448762427102763
ROC-AUC score 0.5189053410893707
Loading Train and Test Data ...
100%|██████████| 2232/2232 [10:33<00:00,  3.52it/s]
100%|██████████| 244/244 [01:07<00:00,  3.63it/s]
Prediction Model SVM
Evaluation of year: 2019
Accuracy 0.5450819672131147
Lag: -64 and cross correlation: 0.17639073009493286
ROC-AUC score 0.5353107344632768
Loading Train and Test Data ...
100%|██████████| 2476/2476 [11:29<00:00,  3.59it/s]
100%|██████████| 251/251 [01:09<00:00,  3.63it/s]
Prediction Model SVM
Evaluation of year: 2020
Accuracy 0.6733067729083665
Lag: 0 and cross correlation: 0.2815450572606813
ROC-AUC score 0.6175571504518873
Loading Train and Test Data ...
100%|██████████| 2727/2727 [12:38<00:00,  3.60it/s]
100%|██████████| 248/248 [01:09<00:00,  3.57it/s]
Prediction Model SVM
Evaluation of year: 2021
Accuracy 0.5967741935483871
Lag: -21 and cross correlation: 0.07890683045023422
ROC-AUC score 0.4966442953020134
```

Results : Combined Text and Price Approach

Performance Metrics of Various Classifiers for NIFTY 50 (Combining Text and Price only) Prediction

	Price representation	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	Open, High, Low, Close, Technical Indicators	BERT	Parallel CNN	0.7003	0.7018	0.3853	0
1	Open, High, Low, Close, Technical Indicators	FinBERT	Parallel CNN	0.7160	0.7105	0.4261	0

Performance Metrics of Various Classifiers for NIFTY NEXT 50 (Combining Text and Price only) Prediction

	Price representation	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	Open, High, Low, Close, Technical Indicators	BERT	Parallel CNN	0.6713	0.6619	0.3300	0
1	Open, High, Low, Close, Technical Indicators	FinBERT	Parallel CNN	0.6885	0.6746	0.4261	0

Performance Metrics of Various Classifiers for NIFTY BANK (Combining Text and Price only) Prediction

	Price representation	Text representation	Classifier	Avg. Accuracy	Avg. ROC-AUC	Avg. NCC	Avg. Lag
0	Open, High, Low, Close, Technical Indicators	BERT	Parallel CNN	0.6895	0.6791	0.4658	0
1	Open, High, Low, Close, Technical Indicators	FinBERT	Parallel CNN	0.6993	0.6812	0.3663	0

This approach integrates news headlines (text) and historical price data for stock prediction using a Parallel CNN model .

Performance Overview:

- Parallel CNN Model:
  - Achieved the highest accuracy and ROC-AUC across all indices.
  - Combined the strengths of FinBERT-based text representations and price data.

Key Findings by Index :

- NIFTY 50 , NIFTY Next 50 and NIFTY Bank :
  - Using FinBERT for text representation yielded the highest accuracy.
  - BERT reduced accuracy by 0.0157 for NIFTY 50 and 0.0098 for NIFTY Bank, with minor drops in ROC-AUC.

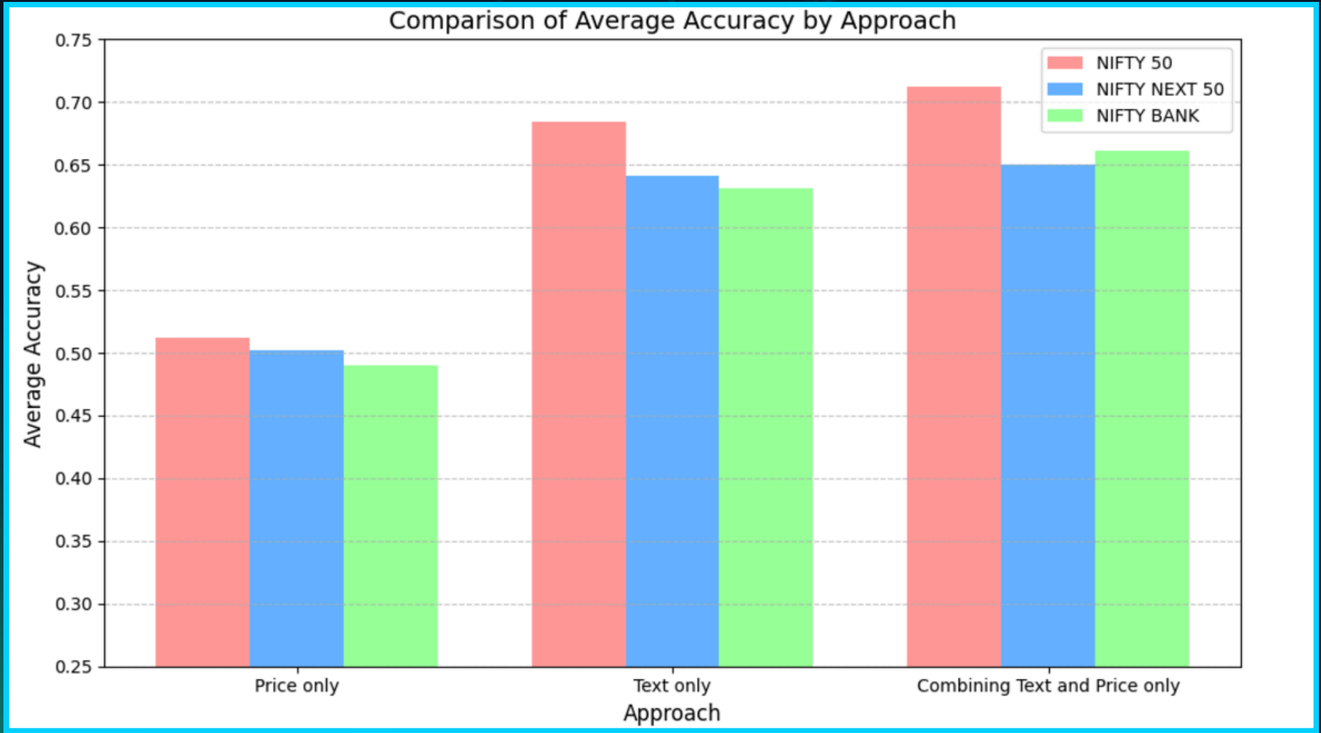
Conclusion:

- Combining text and price data significantly improved performance compared to text-only and price-only methods.
- The Parallel CNN model successfully captured qualitative and quantitative data relationships, providing the most robust predictions.

Comparison Of Best Approaches

Performance Metrics of Various Approaches for NIFTY Indices Prediction

	Approach	Avg. Accuracy for NIFTY 50	Avg. Accuracy for NIFTY NEXT 50	Avg. Accuracy for NIFTY BANK
0	Price only	0.5122	0.5022	0.4901
1	Text only	0.6842 (+25.13%)	0.6412 (+21.67%)	0.6307 (+22.29%)
2	Combining Text and Price only	0.712 (+28.06%)	0.6502 (+22.76%)	0.6613 (+25.88%)



The study concludes that combining text and price data using the Parallel CNN model achieves significantly higher accuracy than using either data source alone. This approach effectively integrates the context of economic events from news articles with patterns in historical prices, validating the hypothesis that news articles published during market hours directly influence stock market movements. Additionally, the use of ROC-AUC and NCC metrics, alongside accuracy, ensures the discovery of more reliable models.

This approach can also be applied to tasks like news recommendation, hate speech detection, fake news classification, and biomedical symptom discovery, leveraging encoder-based representations and CNNs for effective classification.


Thank You