

Article

Not peer-reviewed version

BASIR: Budget-Assisted Sectoral Impact Ranking – A Dataset for Sector Identification and Performance Prediction Using Language Models

[Sohom Ghosh](#)^{*} and Sudip Kumar Naskar

Posted Date: 18 April 2025

doi: 10.20944/preprints202504.1543.v1

Keywords: Financial Texts; Natural Language Processing; Indian Budgets



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

BASIR: Budget-Assisted Sectoral Impact Ranking - A Dataset for Sector Identification and Performance Prediction Using Language Models

Sohom Ghosh * and Sudip Kumar Naskar

Jadavpur University, Kolkata, India

* sohom1ghosh@gmail.com

Abstract: Government fiscal policies, particularly annual union budgets, exert significant influence on financial markets. However, real-time analysis of budgetary impacts on sector-specific equity performance remains methodologically challenging and largely unexplored. This study proposes a framework to systematically identify and rank sectors poised to benefit from India's Union Budget announcements. The framework addresses two core tasks: (1) multi-label classification of excerpts from budget transcripts into 81 predefined economic sectors, and (2) performance ranking of these sectors. Leveraging a comprehensive corpus of Indian Union Budget transcripts from 1947 to 2025, we introduce **BASIR** (**B**udget-**A**ssisted **S**ectoral **I**mpact **R**anking), an annotated dataset mapping excerpts from budgetary transcripts to sectoral impacts. Our architecture incorporates fine-tuned embeddings for sector identification, coupled with language models that rank sectors based on their predicted performances. Our results demonstrate 0.605 F1-score in sector classification, and 0.997 NDCG score in predicting ranks of sectors based on post-budget performances. The methodology enables investors and policymakers to quantify fiscal policy impacts through structured, data-driven insights, addressing critical gaps in manual analysis. The annotated dataset has been released under CC-BY-NC-SA-4.0 license to advance computational economics research.

Keywords: financial texts; natural language processing; Indian Budgets

1. Introduction

In emerging economies, government budget plans greatly influence financial markets.¹ In India, the Union Budget's sector-specific allocations and tax reforms directly influence capital flows, with historical data showing volatility spikes in sectoral indices during budget weeks.² Investors systematically scrutinize budgetary provisions to predict market trajectories.³ Existing works demonstrates significant correlations between budgetary measures and sectoral performance, particularly in consumption-driven industries.⁴ However, current analysis methodologies remain predominantly manual, which is labour-intensive, time consuming, prone to cognitive biases, and may be based on speculation. This study addresses these gaps using a novel computational framework combining transformer-based language models with sectoral performance ranking. We present the tasks in Figure 1.

Our contributions include:

¹ <https://www.ndtv.com/india-news/explained-how-union-budget-influences-stock-market-7517036> (accessed on 16th March, 2025)

² <https://cleartax.in/s/budget-day-market-movement-history-in-india> (accessed on 16th March, 2025)

³ <https://economictimes.com/markets/stocks/news/consumption-over-capex-how-the-budget-impacts-stock-market-investors/articleshow/117853360.cms> (accessed on 16th March, 2025)

⁴ <https://economictimes.com/markets/stocks/news/budget-2025-impact-on-stock-market-which-sectors-stand-to-benefit-or-lose/sector-trends/slideshow/117834567.cms> (accessed on 16th March, 2025)

- **BASIR (Budget-Assisted Sectoral Impact Ranking)** – The first annotated dataset spanning Indian Union Budgets from the year 1947 to 2025, featuring 1,600+ texts from budget transcripts with corresponding sectors labeled. Furthermore, we present 400+ texts with thier corresponding sectoral performance post the day of budget announcement.
- A framework for identifying sectors from budget transcripts and ranking them based on predicted performance.
- Empirical assessment of advanced Large Language Models’ capabilities in predicting performance of sectors based on text excerpts from budget transcripts.

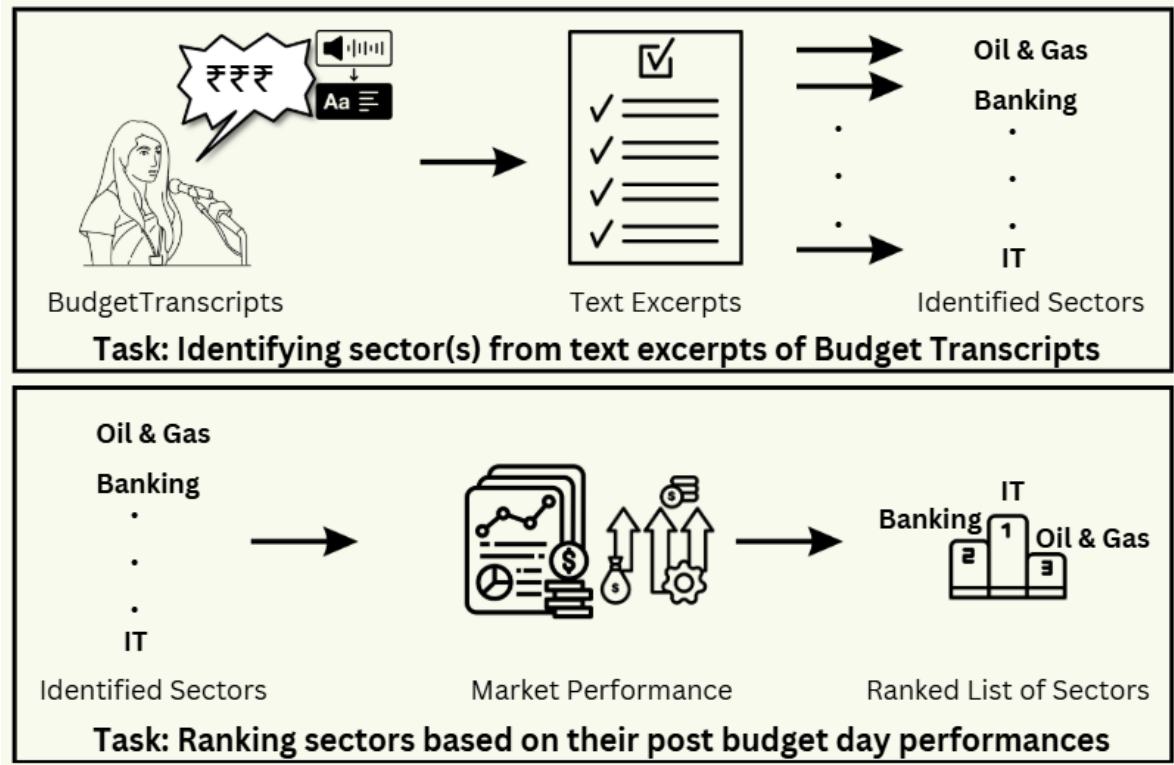


Figure 1. Identifying and Ranking sectors from transcripts of Indian Union Budgets

2. Related Work

The annual Indian Union Budget functions as a crucial instrument for economic policymaking, exerting a direct impact on sectoral growth trajectories and investor sentiment within equity markets [Panwar and Nidugala \(2019\)](#). Research using event study methodology has demonstrated that Cumulative Average Abnormal Returns (CAARs) are significant around budget announcements, indicating that these events contain valuable information for market participants [Kharuri et al. \(2021\)](#); [Manjunatha and Kharuri \(2023\)](#). Our research investigates the empirical evidence of budget-induced stock market reactions across major sectors, with a focus on the transformative role of NLP in this.

Studies [Joshi and Mehta \(2018\)](#); [Martin \(2024\)](#) reveal pronounced sector-specific volatility patterns post-budget announcements, with healthcare, banking, and Information Technology (IT) sectors demonstrating heightened sensitivity to tax reforms and capital allocation decisions. [Mansurali et al. \(2022\)](#) worked on analysing sentiments of tweets relating to Budget 2020. NLP has emerged as a transformative tool in decoding the impacts of fiscal policy on stock markets. Sentiment analysis, a subfield of NLP, is particularly useful in assessing market sentiment and generating trading signals based on prevailing trends [Saxena et al. \(2021\)](#). For instance, advanced NLP models like BERTopic [Grootendorst \(2022\)](#) and RoBERTa [Liu et al. \(2019\)](#) have been employed to analyze the sentiment of the

Reserve Bank of India's monetary policy communications, revealing how different economic topics influence market reactions [Kumar et al. \(2024\)](#).

Most previous studies have focused on post-hoc analyses using historical data, typically conducted after market hours. Our work, however, introduces a predictive approach, innovatively utilizing NLP to automatically detect sectors from budget announcements and rank them according to their predicted performances. This methodological advancement enables the proactive identification of potential market impacts, providing valuable foresight for both investors and policymakers.

3. Problem Statement

This study addresses two sequential challenges in computational fiscal analysis:

1. Multi-Label Sector Classification

Given a budget transcript segment $t \in T$ from India's Union Budget corpus (1947–2025), determine the probabilistic association $P(s_i|t)$ for each sector $s_i \in S$, where $S = \{s_1, \dots, s_{81}\}$ represents formal economic sectors. The task requires overcoming:

- Implicit sector references in policy language (e.g., "Credit access for handloom industries" → Banking, Textile sectors)
- Domain-specific lexical ambiguity (e.g., "digital infrastructure" mapping to both Technology & Utilities sectors)

2. Performance-Aware Sector Ranking

For identified sector set $\hat{S} = \{s_j \mid P(s_j|t) > \tau\}$, develop a model $f : \hat{S} \rightarrow \mathbb{R}^+$ that ranks sectors by expected next day post-announcement returns r_s using text excerpts t related to the sector s_j . Here, τ represents probabilistic threshold.

4. Dataset Construction

Data Collection & Curation

- **Sector-Company Mapping:** We systematically collected a list of sectors and their constituent companies from Screener.in. ⁵
- **Budget Transcripts:** Aggregated 97 Union Budget documents (1947–2025) from India's Ministry of Finance portal ⁶, comprising 1,600+ text excerpts. This also includes the interim budgets.

Annotation Pipeline

1. **Sector Tagging:** For each of the budget transcripts, we prompted DeepSeek [DeepSeek-AI \(2025\)](#) to extract texts and corresponding sector(s) as mentioned in §A.2.1.
2. **Validation:** We manually validated all the outputs.

Market Response Quantification

For sector s in budget day d of a financial year, performance metric $r_{s,d}$ calculated as:

$$r_{s,d} = \frac{1}{|C_s|} \sum_{c \in C_s} \frac{p_{c,d+1}^{\text{open}} - p_{c,d}^{\text{open}}}{p_{c,d}^{\text{open}}}$$

where C_s denotes constituent companies of sectors, with historical data sourced from yahoo finance. ⁷ $p_{c,d}^{\text{open}}$ denotes the opening price of company c on day d . Finally, we ranked the sectors in decreasing order of their performances. More details about the data is presented in Table 1. Data till the year 2019

⁵ <https://www.screener.in/explore/> (accessed on 17th March, 2025)

⁶ <https://www.indiabudget.gov.in/bspeech.php> (accessed on 17th March, 2025)

⁷ <https://finance.yahoo.com/> (accessed on 17th March, 2025)

was used for training, data spanning 2020 to 2023 was allocated for validation, and 2024 data was reserved for testing.

Table 1. Dataset Statistics

Metric	Budget Transcripts	Sector Identification	Sector Ranking
Total Entries	97	1,671	429
Temporal Span	1947–2025	1947–2025	1997–2025

5. Experiments & Results

This study involved two primary experimental components. Firstly, we employed a methodology to identify specific sectors from excerpts of budget transcripts. Secondly, we developed a framework to rank these identified sectors based on their performance, thereby providing a comprehensive analysis of sectoral impacts.

5.1. Identifying Sectors from Excerpts of Budget Transcripts

The task of identifying sectors from budget excerpts was approached as a multi-class classification problem. We implemented and evaluated several methodologies to address this challenge.

Initially, we employed semantic similarity (STS) based on Nomic embeddings [Nussbaum et al. \(2024\)](#) to identify sectors from given text excerpts. To enhance performance, we subsequently fine-tuned these embeddings to optimize the vector space representation, such that sectors relevant to a particular excerpt were positioned closer together, while unrelated sectors were distanced. Additionally, we fine-tuned pre-trained language models, specifically BERT [Devlin et al. \(2019\)](#), and RoBERTa [Liu et al. \(2019\)](#), for the classification of budget excerpts into appropriate sectors.

The performance metrics for the various models are presented in Table 2. Our analysis reveals that the STS model with fine-tuned embeddings, and τ equals to 0.5 demonstrated superior performance in terms of both Macro (M) and Weighted (W) F1 scores. This suggests that the fine-tuned embedding approach effectively captures the nuanced relationships between budget language and sectoral classifications. Conversely, the BERT model exhibited the highest Micro (m) F1 score, indicating its strength in correctly classifying the most frequent sector categories.

Table 2. Results of Multi-Label Sector Classification

	F1 (M)	F1 (m)	F1 (w)
STS (base)	0.159	0.176	0.345
STS (fine-tune)	0.291	0.478	0.605
BERT	0.179	0.489	0.425
RoBERTa	0.075	0.274	0.192

5.2. Ranking Sectors Based on Their Performance

To rank sectors based on their performance, we developed and evaluated four distinct architectural approaches.

Our initial approach involved transforming sector performance data into a binary classification task, determining whether a given sector would experience an upward or downward movement based on the text excerpts related to it. Using this framework, we fine-tuned three encoder-based (Enc) models: BERT [Devlin et al. \(2019\)](#), RoBERTa [Liu et al. \(2019\)](#), and DeBERTa [He et al. \(2020\)](#) for classification purposes. The predicted probabilities from these models were then utilized to generate sector rankings.

Building upon this classification approach, we subsequently fine-tuned the same models for regression analysis. This allowed us to predict the actual performance metrics for each sector with

greater precision. The sectors were then ranked according to these predicted performance values, providing a more nuanced assessment of relative sectoral strength.

Following our encoder-based approaches, we implemented feature-based models utilizing Nomic embeddings [Nussbaum et al. \(2024\)](#) (Emd) extracted from sector-related text excerpts. For binary classification, we trained several machine learning algorithms including logistic regression, random forest, and XGBoost [Chen and Guestrin \(2016\)](#). These models were tasked with predicting whether sectors would experience positive or negative performance.

In parallel, we developed regression models using linear regression, random forest, and XGBoost algorithms to predict the actual performance metrics of each sector. The ranking methodology remained consistent with our previous approaches, wherein sectors were ordered based on their predicted performance values. Additionally, we trained an XGBoost model specifically optimized with a learning-to-rank objective to directly produce sector rankings.

In our final experimental approach, we leveraged state-of-the-art large language models (LLMs) to estimate sector performance based on budget text excerpts. Specifically, we employed three advanced LLMs: Gemma-3 27B [Team \(2025\)](#), DeepSeek V3 [DeepSeek-AI et al. \(2025\)](#), and Llama 3.3 70B [Touvron et al. \(2023\)](#). These models were prompted to analyze the sector-relevant text excerpts and estimate the expected performance metrics for each sector. The resulting performance estimates were then utilized to generate sector rankings. More details of the prompts are provided in [§A.2.2](#).

Table 3 presents the comparative performance metrics for these these architectural approaches. Notably, the BERT model trained for classification exhibited superior performance in terms of Normalized Discounted Cumulative Gain (NDCG), suggesting that smaller models are more effective when we have lesser number of instances to train. The performance of the LLMs is comparable to that of the other approaches.

Table 3. Sector Ranking Results

Model	Type	NDCG
BERT	Enc Clasifier	0.997
RoBERTa	Enc Clasifier	0.994
DeBERTa	Enc Clasifier	0.996
BERT	Enc Regressor	0.995
RoBERTa	Enc Regressor	0.995
DeBERTa	Enc Regressor	0.995
Logistic	Emd + Classifier	0.996
Random Forest	Emd + Classifier	0.996
XG-Boost	Emd + Classifier	0.994
Linear	Emd + Regressor	0.995
Random Forest	Emd + Regressor	0.996
XG-Boost	Emd + Regressor	0.994
XG-Boost	Learning to Rank	0.994
Gemma-3 27B	Zero Shot	0.994
DeepSeek V3	Zero Shot	0.993
Llama 3.3 70B	Zero Shot	0.994

6. Conclusion

This study presents a comprehensive framework for the detection and performance-based ranking of sectors from Indian Union Budget transcripts. Our findings demonstrate that fine-tuned Nomic-based embeddings provide superior performance in identifying sectors from textual excerpts, capturing the nuanced relationships between budget language and sectoral classifications. Concurrently, the BERT-based model fine-tuned for classification emerged as the most effective approach for ranking sectors based on predicted performance, surpassing other methodologies.

The framework developed in this research offers valuable insights for investors, and financial analysts seeking to understand the immediate market implications of budget announcements. By automating the process of sector identification and performance prediction, our approach enables more timely and informed decision-making.

Future research directions include extending this framework to recommend specific stocks within the identified sectors, potentially offering more granular investment guidance. Additionally, developing capabilities to capture real-time price movements following budget announcements would enhance the practical applicability of this work.

Limitations

Despite our methodological contributions, several limitations warrant acknowledgment.

First, our annotation approach emphasized precision over recall in sector identification. The DeepSeek language model may have overlooked subtler budget-sector relationships, particularly when policy implications were implicit rather than explicit. Our validation protocol—focusing exclusively on LLM-detected relationships—potentially reinforces this detection bias, creating systematic blind spots in the dataset. Consequently, fiscal impacts on certain sectors may be underrepresented in our analysis.

Second, temporal coverage presents significant constraints. Market performance data availability beginning only from 1997 excluded 50 years of budget documents (1947-1996) from complete analysis. This limitation is particularly significant when analyzing long-term policy impacts and historical shifts in sector prioritization. Additionally, inconsistent market data across sectors forced the exclusion of certain sector-period combinations, introducing potential selection bias. These gaps disproportionately affected newly formalized sectors and those with limited public listings.

Third, our performance metric isolates budget effects without controlling for confounding variables. Macroeconomic factors (monetary policy adjustments, global market movements), sector-specific events (regulatory changes, technological disruptions), and concurrent corporate announcements likely influence post-budget market movements. The absence of a comprehensive control framework limits causal interpretations of budget-performance relationships.

Future research should address these limitations through multi-source validation, synthetic data generation for pre-1997 periods, and development of counterfactual models that control for non-budgetary market influences.

Appendix A

Appendix A.1 Industries

List of industries are as follows: ['Aerospace & Defence', 'Agro Chemicals', 'Air Transport Service', 'Alcoholic Beverages', 'Auto Ancillaries', 'Automobile', 'Banks', 'Bearings', 'Cables', 'Capital Goods - Electrical Equipment', 'Capital Goods-Non Electrical Equipment', 'Castings', 'Forgings & Fasteners', 'Cement', 'Cement - Products', 'Ceramic Products', 'Chemicals', 'Computer Education', 'Construction', 'Consumer Durables', 'Credit Rating Agencies', 'Crude Oil & Natural Gas', 'Diamond, Gems and Jewellery', 'Diversified', 'Dry cells', 'E-Commerce/App based Aggregator', 'Edible Oil', 'Education', 'Electronics', 'Engineering', 'Entertainment', 'Ferro Alloys', 'Fertilizers', 'Finance', 'Financial Services', 'FMCG', 'Gas Distribution', 'Glass & Glass Products', 'Healthcare', 'Hotels & Restaurants', 'Infrastructure Developers & Operators', 'Infrastructure Investment Trusts', 'Insurance', 'IT - Hardware', 'IT - Software', 'Leather', 'Logistics', 'Marine Port & Services', 'Media - Print/Television/Radio', 'Mining & Mineral products', 'Miscellaneous', 'Non Ferrous Metals', 'Oil Drill/Allied', 'Packaging', 'Paints/Varnish', 'Paper', 'Petrochemicals', 'Pharmaceuticals', 'Plantation & Plantation Products', 'Plastic products', 'Plywood Boards/Laminates', 'Power Generation & Distribution', 'Power Infrastructure', 'Printing & Stationery', 'Quick Service Restaurant']

, 'Railways', 'Readymade Garments/ Apparels', 'Real Estate Investment Trusts', 'Realty', 'Refineries', 'Refractories', 'Retail', 'Ship Building', 'Shipping', 'Steel', 'Stock/ Commodity Brokers', 'Sugar', 'Telecomm Equipment & Infra Services', 'Telecomm-Service', 'Textiles', 'Tobacco Products', 'Trading', 'Tyres']

Appendix A.2 Prompts

Appendix A.2.1 Text Extraction and Sector Identification

You are provided with the budget of India below. From this budget only pick up text segments relevant to the given list of industries. List of industries: <list of industries> Your output should be a json file having 2 keys: 'text_segment' and 'industry'. The value corresponding to 'text_segment' would be the extract text segment extracted from the budget. The value of 'industry' should be the corresponding list of industries from the given list that the text segment is related to. Return only the segments having any relation with the given list of industries. One text segment can be related to multiple industries.

Text context from Budget: <Budget Transcript of a given year>

Appendix A.2.2 Sectorwise Performance Prediction

You are a financial expert with extensive experience of analysing Indian Budgets. Given a sector and an excerpts related to the sector from a budget speech, estimate the performance of the sector. Your output should be just a real number between -1 to 1. Don't reply anything else. Sector: <name of sector>, Excerpt: <text excerpts related to the given sector>

Appendix A.3 Reproducibility

The codes and the datasets can be accessed from https://huggingface.co/datasets/sohomghosh/BASIR_Budget_Assisted_Sectoral_Impact_Ranking/tree/main/

References

- Chen, Tianqi and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 785–794. ACM. <https://doi.org/10.1145/2939672.2939785>.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.
- DeepSeek-AI, Aixin Liu, and Bei Feng et al. 2025. Deepseek-v3 technical report.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019, June. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.
- Grootendorst, Maarten. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Joshi, Mrunal and Rucha Mehta. 2018. Impact of union budget on stock market. *Contemporary Issues in Marketing and Finance* 1, 29–45.
- Kharuri, Zahid Hassan, T Manjunatha, and V Rajesh Kumar. 2021. Stock price reactions to budget announcement in indian capital market. *International Journal of Science and Management Studies* 4(6), 59–69.
- Kumar, Rohit, Sourabh Bikas Paul, and Nikita Singh. 2024. Words that move markets-quantifying the impact of rbi's monetary policy communications on indian financial market. *arXiv preprint arXiv:2411.04808*.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

- Manjunatha, T and Zahid Hassan Kharuri. 2023. Effects of budget announcement on stock prices in the indian context. *Asian Journal of Management* 14(1), 57–64.
- Mansurali, A, P Mary Jayanthi, R Swamynathan, and Tanupriya Choudhury. 2022. Social listening on budget—a study of sentimental analysis and prediction of sentiments using text analytics & predictive algorithms. In *Machine Intelligence and Data Science Applications: Proceedings of MIDAS 2021*, pp. 879–892. Springer.
- Martin, Geo. 2024. Analyzing the impact of the union budget on sectoral indices in the national stock exchange (nse).
- Nussbaum, Zach, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder.
- Panwar, Vivek and Ganesh Kumar Nidugala. 2019. Impact of budget and gdp announcements on indian stock market. *Finance India* 33(4), 929–946.
- Saxena, Anshul, Vandana Vijay Bhagat, and Amrita Tamang. 2021. Stock market trend analysis on indian financial news headlines with natural language processing. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–5. IEEE.
- Team, Gemma. 2025. Gemma 3.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.