Measuring Progress in Sustainable Development Goals (SDGs): An Application of Natural Language Processing (NLP) on Budget Documents

by Rajni Dahiya and Shashi Kant ^

This article examines the progress in terms of focus on SDGs across the Union Government and select States in India by looking at the budget documents. Using topic modelling, each theme is assigned to one or more SDGs based on multi-label classification, to determine whether there are systematic or structural differences in policies on SDGs as outlined in budget documents. Results show that the combined focus on SDGs by the Union Government and select States has increased in 2023 compared to 2012. It is observed that the emphasis on different SDGs varies over time depending on the changing circumstances and there is considerable heterogeneity among the Union Government and select States with reference to their focus on the individual SDGs.

Introduction

SDGs are a set of 17 ambitious global goals tackling critical challenges *i.e.,* poverty, inequality, climate change, and environmental degradation which emerged from the Rio+20 conference in 2012 and were formally adopted by all United Nations (UN) member countries in 2015 (Table A1 in Annex provides a detailed description of each of these goals).[1] SDGs are successors to the millennium development goals (MDGs) of 2000, but they go further by building upon previous successes and lessons learned. They offer a framework for collaborative action towards a more sustainable and equitable world by 2030.

Comprehending the advancements achieved for each goal under SDG empowers policymakers, organisations, and individuals to take purposeful, outcome-driven actions. Taking a step in this direction, the first edition of the SDG India Index was launched in December 2018, incorporating 62 indicators from 39 targets across 13 SDGs.[2] *NITI Aayog* releases the SDG India Index, which assesses the progress made by all States and Union Territories (UTs) towards achieving the SDGs by utilising the latest data provided by ministries and departments. The index is based on the National Indicator Framework developed by the Ministry of Statistics and Programme Implementation (MoSPI) in consultation with *NITI Aayog*. Normalised scores are calculated for each State/UT and the composite index of a goal is computed using an arithmetic mean of the normalised values of all indicators related to the goal. Each indicator within each State/UT is assigned equal weight.

This article offers a novel approach to measure progress in SDGs by leveraging advancements in NLP. This approach aims to bridge the data gap in emerging market economies and construct time series in case backward data is not available (Conforti *et al.,* 2020). Additionally, it may help in automated text labelling of vast documents from diverse sources overcoming the limitations of manual labelling, which may be time-consuming and prone to bias (Guisiano *et al.,* 2022). The methodology in this study uses NLP techniques to extract information from budget documents and deploys a large language model (LLM) to study whether there are systematic or structural differences in policies on SDGs as outlined in budget documents. In the digital age, NLP algorithms,

^ The authors are from the Department of Economic and Policy Research (DEPR). We sincerely thank Dr G.V.Nadhanael for his insightful comments which have significantly improved the quality of the article. The views expressed in this article are those of the authors and do not represent the views of the RBI.

[1] United Nations SDGs: https://sdgs.un.org/goals.

[2] SDG India Index Methodology, 2018: https://sdgindiaindex.niti.gov.in/assets/Files/SDG%20India%20Index%202018%20Methodology.docx.

especially transformer models, are instrumental in making sense of vast unstructured text data, helping machines process information efficiently (Vaswani *et al.*, 2017). The basics of NLP are presented in the Annex.

The available studies on progress in SDGs in India have largely focused on challenges, implementation, and state-wise comparison of performance (Mohandas, 2018; Panda *et al.*, 2018); While several studies have explored the intersection of SDGs and NLP in other country contexts such as Amel-Zadeh *et al.*, (2021) and Matsui *et al.*, (2022), this article contributes to the literature by providing a theme-based SDG classification within large textual datasets. Also, by using a blend of NLP techniques and network analysis to chart relationships among SDGs, this article explores interconnectedness across them. This co-occurrence analysis reveals important thematic connections and identifies potential synergies between the SDGs (Smith *et al.*, 2021; Zhou *et al.*, 2022).

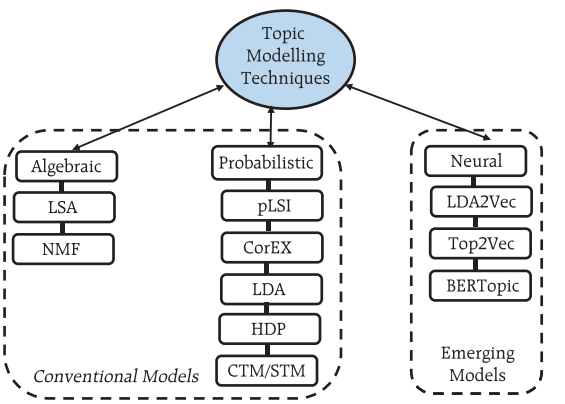
Overall, NLP techniques can complement quantitative assessments like the SDG India Index. By analysing budget documents of the Union Government as well as select State Governments based on availability, the article deploys NLP to unveil the government's overall narrative and messaging surrounding SDGs. It delves into government priorities and narratives through thematic analysis, identifying recurring themes and topics in budget documents, and offering a deeper understanding beyond mere numbers. It also allows for comparative analysis, offering insights into how different government entities tackle SDGs and track trends in the government's communication focus over time. Accordingly, the remainder of the article is organised as follows: Section II discusses topic modelling; Section III delves into data and methodology; Section IV details the results, and Section V concludes the analysis.

II. NLP and Topic Modelling: Some Preliminaries

Topic modelling in NLP is an unsupervised learning approach for organising text documents based on their underlying semantic structure. From an algorithmic perspective, topic modelling techniques can be broadly classified into three main categories: algebraic, probabilistic, and neural models (Vayansky *et al.*, 2020; Abdelrazek *et al.*, 2023). The first two fall under traditional statistics-based methods, while the third represents the more contemporary approach of employing artificial neural networks in NLP (Chart 1).

Among conventional models, LDA has been the dominant choice for decades (Blei *et al.*, 2003). However, many studies have uncritically applied LDA, neglecting justification for their topic modelling method selection. A significant drawback of LDA is its dependence on the bag-of-words (BoW) approach, neglecting semantic relationships among words (Chen *et al.*, 2023). In general, conventional topic modelling techniques such as LDA necessitate intricate corpus pre-processing, meticulous parameter selection (*e.g.*, determining the number of topics), proper model evaluation, and interpretation of generated topics relying on both common sense

Chart 1: Categories of Topic Models



Note: In traditional models, algebraic models comprise latent semantic analysis (LSA), and non-negative matrix factorization (NMF).The probabilistic models encompass probabilistic latent semantic indexing (pLSI), anchored correlation explanation (CorEx) [Gallagher *et al.*, 2017], latent dirichlet allocation (LDA) [Blei *et al.*, 2003], along with numerous extensions and variants of LDA, including hierarchical dirichlet process (HDP), correlated topic model (CTM), and structural topic model (STM).
Sources: Chen, *et al.*, (2023); and Authors' illustration.

and domain knowledge. Recently developed neural models have gained significant popularity since 2016. Examples within the neural category encompass LDA2Vec (Moody, 2016), SBM (Gerlach *et al.*, 2018), deep LDA (Bhat *et al.*, 2020), Top2Vec (Angelov, 2020), and BERTopic (Grootendorst, 2022). This trend aligns with the exponential progress of deep learning in recent years.

III. Data and Methodology

This study utilises budget speeches from select Indian States (Andhra Pradesh, Karnataka, Kerala, Odisha, Tamil Nadu, and West Bengal) and the Union Government, spanning from 2012 to 2023. As shown in Chart 2, after text pre-processing of the respective budget speeches, BERTopic was employed for generating distinct themes as it outperformed traditional models *i.e.*, NMF, LSA, and LDA (Egger *et al.*, 2022; Chen *et al.*, 2023).[3] These themes were further classified into specific SDGs using Hugging Face's pre-trained SDG classifier model.[4] It is important to note that this study focuses on the first 15 SDGs, as the pre-trained model is fine-tuned based on publicly available data from the OSDG community (OSDG, 2022) which classifies the data in the first 15 SDGs.

In order to measure the evolution of emphasis on SDG goals over time, a key tool in NLP known as TF-IDF[5], is used. The intuition behind TF-IDF is that a term's importance is inversely related to its frequency across documents. This score combines two factors: term frequency (TF) which measures how often a term appears in a document, and inverse document frequency (IDF) which penalises common terms across the entire corpus.

TF-IDF scores are calculated by multiplying TF and IDF.

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

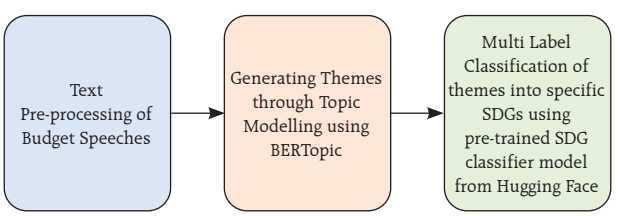
A detailed explanation about TF-IDF is provided in the Annex.

IV. Results

First, the overall evolution of communication strategy in budget documents is analysed. It is found that budget speeches length marginally decreased, in terms of word counts, for the Union Government, Kerala and Karnataka, while it increased for other States in 2023-24 compared to 2012-13 (Chart 3).

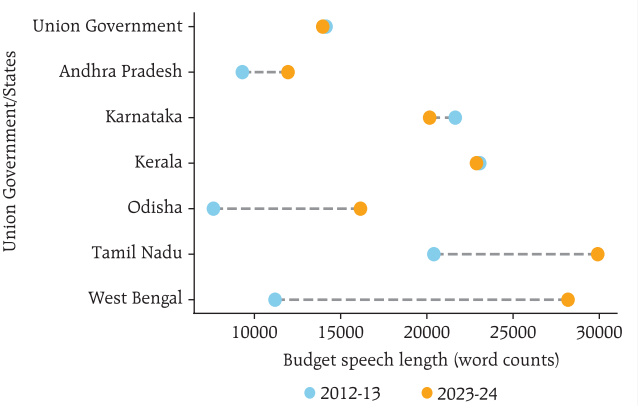
Looking beyond number of words, an analysis of unique number of topics mentioned in the budgets were attempted. Through topic modelling, amongst

Chart 2: Theme-Based SDG Classification Pipeline



Source: Authors' illustration.

Chart 3: Length of Budget Speech



Notes: 1. Tamil Nadu and Odisha had separate budget speeches dedicated to agriculture in 2023-24, similar to Karnataka in 2012-13. The budget speeches have been merged together for analysis.
2. The chart represents the initial word counts of the budget speeches before data pre-processing.

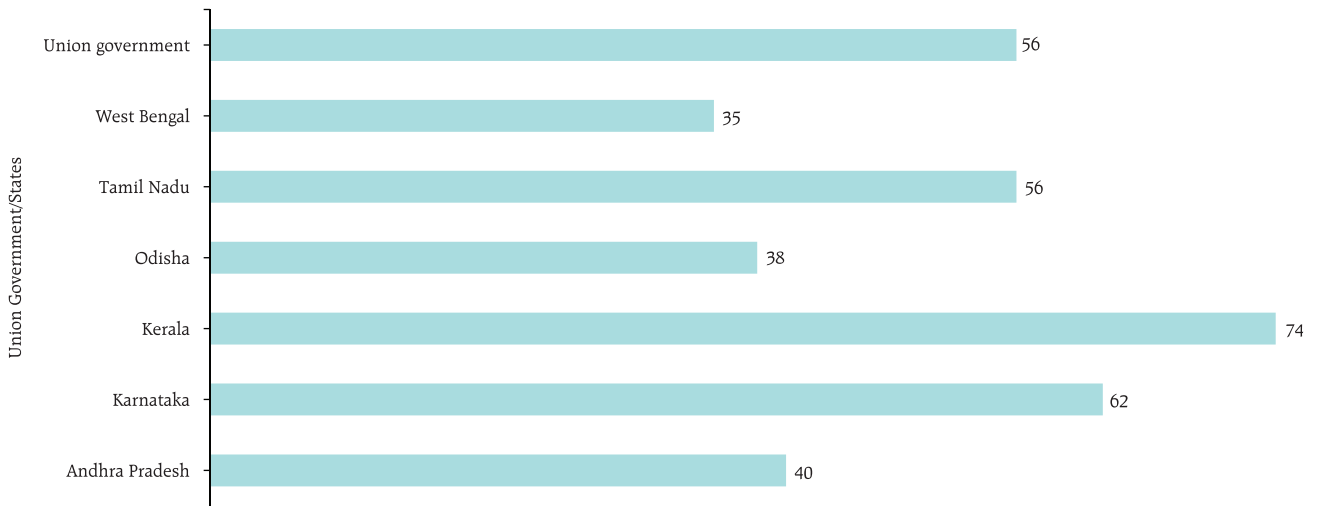
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

[3] BERTopic is successful in capturing context and word semantics due to its use of bidirectional encoder representations transformer (BERT) word embeddings (Devlin *et al.*, 2019; Amin *et al.*, 2022).

[4] Pre-trained SDG classifier model is available at: https://huggingface.co/jonas/sdg_classifier_osdg.

[5] Term Frequency-Inverse Document Frequency (TF-IDF).

Chart 4: Number of Topics



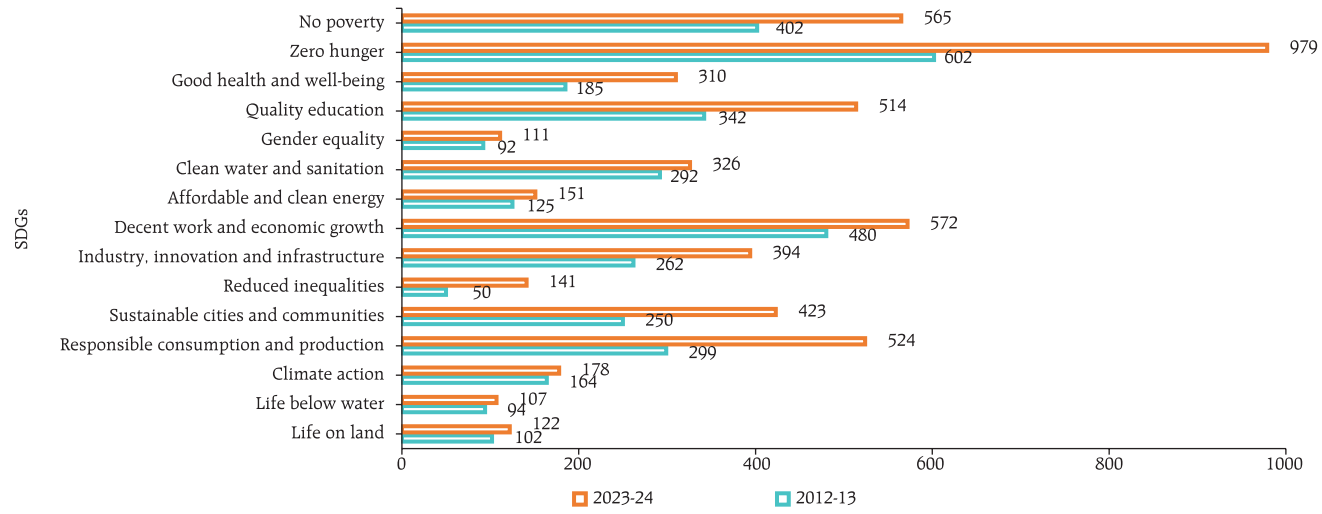
Note: Theme (-)1 has been removed as it usually contains either generic topics or a combination of irrelevant words hence is treated as an outlier (Grootendorst, 2022), and a few additional themes were also removed during text cleaning as they were not aligned with SDGs. Total number of budget documents analysed were 84.
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

the sample states, Kerala's budget speeches emerged as the most diverse in SDG themes, followed by Karnataka, the Union Government, and Tamil Nadu (Chart 4).

IV.1 Transition in SDGs Focus over Time

Results show that the combined focus by the Union Government and select States has increased in 2023-24 compared to 2012-13 on all fifteen SDGs, indicating their stronger commitment towards achieving these goals by 2030. In particular, the focus on zero hunger (SDG 2), good health and well-being (SDG 3), reduced inequalities (SDG 10), sustainable cities and communities (SDG 11), and responsible consumption and production (SDG 12) has witnessed a remarkable increase in 2023-24 compared to 2012-13 (Chart 5).

Chart 5: SDG Focus Transition (2023-24 vs 2012-13)



Note: Frequency, derived from c-TF-IDF (details in Annex), is aggregated for both the Union Government and select States.
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

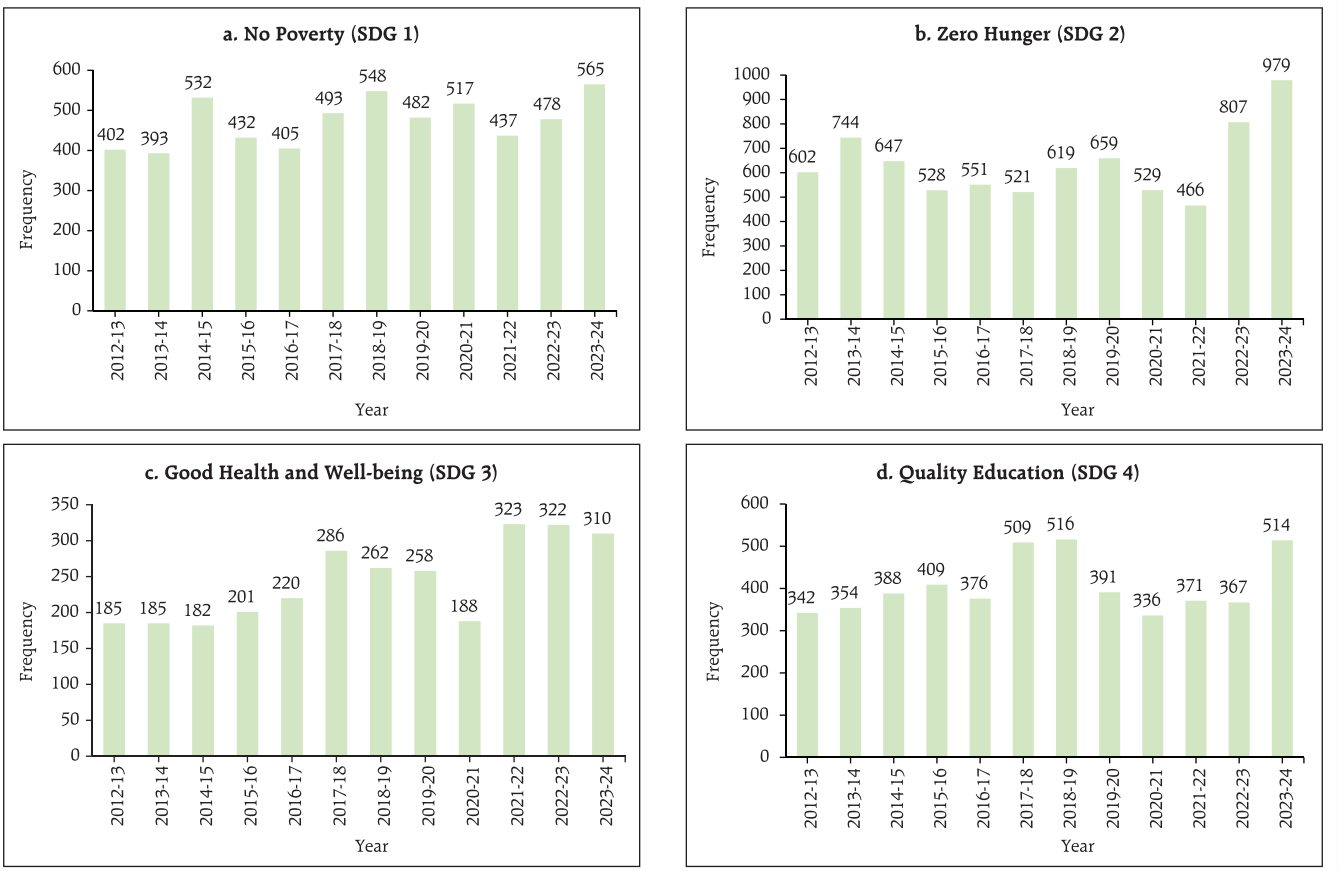
IV.2 Trend in SDGs

To delve deeper into the trends within the key SDGs across 2012-13 to 2023-24, the combined focus by the Union Government and select States on the first four SDGs was analysed. This analysis reveals variation in focus over the years. In 2023-24, both no poverty (SDG 1) and zero hunger (SDG 2) reached their highest focus (Chart 6a and 6b). The COVID-19 pandemic significantly impacted the focus on good health and well-being (SDG 3), with a steep rise in 2021-22, reaching its highest level in the last nine years (Chart 6c). For quality education (SDG 4), both 2018-19 and 2023-24 saw significant reforms in the education sector. Key schemes *i.e.*, *Samagra Shiksha Abhiyan*[6] (2018) and the National Education Policy (2023) mark these years as pivotal for education development in India (Chart 6d).

IV.3 Mapping the Interconnectedness of the SDGs: A Co-Occurrence Analysis

SDGs are not mutually exclusive and targeting a particular SDG may help achieve other linked goals as well. Analysing these co-occurrences renders valuable insights into the interconnectedness

Chart 6: Trend in Key SDGs Combined across the Union Government and States

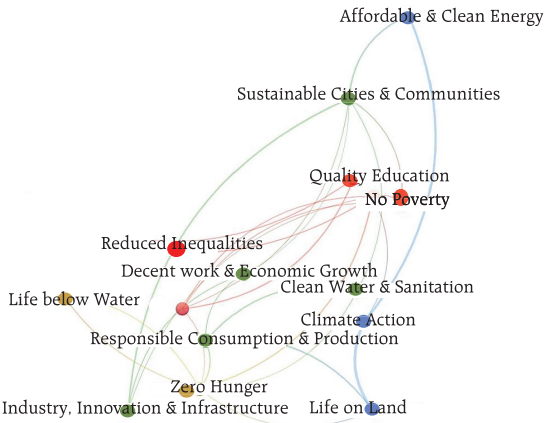


Notes: 1. The frequency metric used here is the c-tf-idf score.
2. While the budget speeches address all the fifteen SDGs set by the United Nations, this paper shows the trend only in the first four SDGs to demonstrate the analysis approach. Similar visualisations can be created to explore trends in the remaining SDGs.
3. In election years, for the Union Government and select States, the budget speeches analysed are the final speeches, not the interim speeches.

Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

[6] The *Samagra Shiksha* scheme is an integrated scheme for school education covering the entire gamut from pre-school to class XII. The scheme treats school education as a continuum and is in accordance with SDG 4. About the scheme: https://dsel.education.gov.in/scheme/samagra-shiksha.

Chart 7: Co-occurrence of SDGs based on Overlapping Themes



Note: The thickness of the lines connecting SDGs in the chart indicates the number of shared themes between them, with thicker lines representing more themes.
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

of the SDGs which helps in the identification of integrated policy approaches that address multiple goals simultaneously. The analysis of co-occurrence of SDGs based on overlapping themes is presented in Chart 7. The theme related to renewable energy was tagged with affordable and clean energy (SDG 7), responsible consumption and production (SDG 12), and climate action (SDG 13) because renewable energy directly addresses the goal of providing clean and affordable energy, promoting sustainable consumption practices and mitigating climate change risk. Similarly, a theme focused on forests and biodiversity was tagged with climate action (SDG 13) and life on land (SDG 15). This reflects the crucial role forests play in sustaining life on land and mitigating the impact of climate change.

IV.4 Policy Alignment between the Union Government and States

Amel-Zadeh *et al.* (2021) employed NLP techniques to pinpoint companies that aligned with the UN SDGs through the analysis of the text in their sustainability disclosures. The same principles may be extended to measure alignment of the Union Government and States with various SDGs. *A priori*, policy focus of the Centre and the States is expected to be different and also could vary over time. The constitution divides subjects of governance into three verticals *i.e.*, Union, States, and Concurrent lists. This gives the national and sub-national governments the freedom and space to act on the subjects in their jurisdiction with some degree of overlap in the concurrent list where the two can cooperate and complement each other's policies. The States can make policies fine-tuned to suit their local conditions. The hypothesis gets reaffirmed when this study reduces the dimensionality of the 15 SDGs to a 2-dimensional plane using two principal components. Each budget document can be represented by a 15x1 vector in terms of the focus it has on various SDGs. To visualise the orientation of policies, the dimensionality is reduced from a 15-D space to a 2-D space by plotting the first two principal components which together represent 47 per cent of the total variance (Table 1).

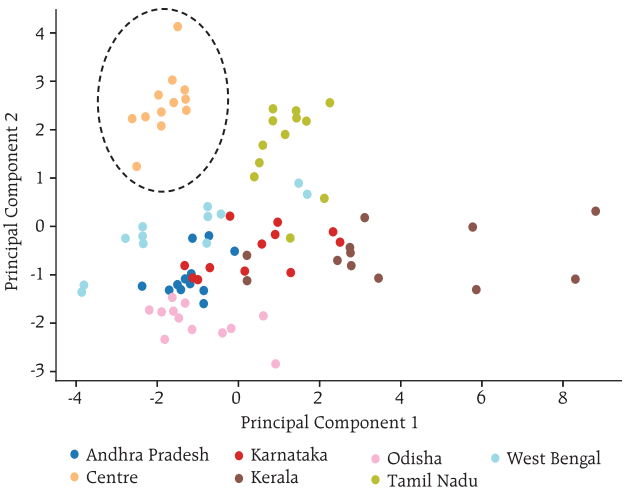
The visual representation depicts that the topics covered in budget documents of Union Government (cluster highlighted by black dots), and States are segregated from each other (Chart 8). While policies of some States show less variance across the years, others display considerable diversity in emphasis

Table 1: Cumulative Variation Explained by Successive Principal Component

PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
0.32	0.47	0.60	0.67	0.74	0.80	0.84	0.88	0.91	0.93
PC 11	**PC 12**	**PC 13**	**PC 14**	**PC 15**					
0.96	0.97	0.98	0.99	1.00					

Note: Values are rounded off to two decimal points.
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

Chart 8: Principal Component Analysis



Note: Each State/Union Government is represented by 12 circles, which indicate the years from 2012-13 to 2023-24.
Sources: Union Government and State Budget Speeches 2012-13 to 2023-24; and Authors' calculations.

on SDGs. The Union Government exhibits the least variance across the years compared to the States (Table 2). These differences suggest that the State level policies might be devised keeping in view their local situations.

V. Conclusion

NLP has become an invaluable tool for enhancing data analysis and is increasingly being applied in policymaking. This article proposes a novel use of NLP to measure progress towards SDGs in the Indian context, complementing existing quantitative approaches. The analysis undertaken for the Union Government and select States indicates that their combined focus on SDGs was higher in 2023-24 as compared to 2012-13 pointing towards a stronger commitment towards achieving these goals by 2030. The focus on various SDGs has varied over time depending on the changing circumstances such as an increase in focus on health during the COVID-19 pandemic. It is also found that SDGs are not mutually exclusive and exhibit significant interconnectedness between them. Therefore, targeting a particular SDG may help achieve other linked goals too. Furthermore, this approach helps to check the heterogeneity in focus on SDGs among Union Government and select States. Going beyond the traditional quantitative measures, the analysis of relative focus on SDGs presented in this article provides important policy insights into the evolution, interconnectedness, and relative focus over time.

Table 2: Clusters Size in SDG Representation of Budget Documents

Region	Average Cluster Size
Kerala	3.50
Karnataka	2.84
Tamil Nadu	2.33
West Bengal	2.26
Andhra Pradesh	2.06
Odisha	1.93
Union Government	1.78

Note: The average cluster size is calculated using Euclidean distance of each data point from the centroid of each cluster in Chart 8.
Source: Authors' calculations.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2023). Topic Modeling Algorithms and Applications: A Survey. *Information Systems, 112, 102131.*

Amel-Zadeh, A., Chen, M., Mussalli, G., and Weinberg, M. (2021). NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals. *Social Science Research Network (SSRN).*

Amin, A., Hassan, S., Alm, C., and Huenerfauth, M. (2022). Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users. *ResearchGate Publication.* https://doi.org/10.13140/RG.2.2.28272.33289

Anastasopoulos, L., Moldogazi, T., and Scott, T. (2017). Computational Text Analysis for Public Management Research. *SSRN Electronic Journal, doi:10.2139/ssrn.3269520.*

Angelov, D. (2020). Top2vec: Distributed Representations of Topics. *arXiv 2020, arXiv:2008.09470.*

Antonellis, I., and Gallopoulos, E. (2006). Exploring Term-Document Matrices from Matrix Models in Text Mining. arXiv: cs/0602076.

Bhat, M. R., Kundroo, M. A., Tarray, T. A., and Agarwal, B. (2020). Deep LDA: A New Way to Topic Model. *Journal of Information Optimization Sciences, 41, 823–834.*

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3, 993–1022.*

Camacho-Collados, J., and Pilehvar, M. T. (2017). On the Role of Text Pre-processing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *arXiv, 1707.01780.*

Cavnar, W. B., and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. *Proceedings of SDAIR-94, 161–175.*

Chen, W., Rabhi, F., Liao, W., and Al-Qudah, I. (2023). Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics, 12(12), 2605.*

Christian, H., Agus, M., and Suhartono, D. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications, 7, 285.*

Conforti, C., Hirmer, S., Morgan, D., Basaldella, M., and Ben Or, Y. (2020). Natural Language Processing for Achieving Sustainable Development: The Case of Neural Labelling to Enhance Community Profiling. *ACL Anthology.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics.*

Egger, R., and Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology, 7.*

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Transactions of the Association for Computational Linguistics*, 5, 529–542.

Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018). A Network Approach to Topic Models. *Science Advances, 4, eaaq1360.*

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv 2022, arXiv:2203.05794.*

Guisiano, J. E., Chiky, R., and De Mello, J. (2022). SDG-Meter: A Deep Learning Based Tool for Automatic Text Classification of the Sustainable Development Goals. *United Nations Environment Program, Paris, France.*

Hachaj, T., and Ogiela, M. R. (2018). What Can Be Learned from Bigrams Analysis of Messages in Social Network? In *2018 11th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)* [pp. 1-4]. Beijing, China. doi:10.1109/CISP-BMEI.2018.8633108.

Khyani, D., and B S, S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology, 22*, 350-357.

Kurniasih, A., and Manik, L. P. (2022). On the Role of Text Pre-processing in BERT Embedding-based DNNs for Classifying Informal Texts. *International Journal of Advanced Computer Science and Applications (IJACSA), 13(6).*

Matsui, T., Suzuki, K., Ando, K., Kitai, Y., Haga, C., Masuhara, N., and Kawakubo, S. (2022). A Natural Language Processing Model for Supporting Sustainable Development Goals: Translating Semantics, Visualizing Nexus, and Connecting Stakeholders. *Sustain Sci, 17, 969–985.*

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to make Lda2vec. *arXiv, arXiv:1605.02019.*

Mohandas, P. (2018). Sustainable Development Goals (SDGs) - Challenges for India. *Indian Journal of Public Health Research & Development, 9*(1), 1. https://doi.org/10.5958/0976-5506.2018.00172.9

OSDG, UNDP IICPSD SDG AI Lab, and PPMI (2022). OSDG Community Dataset. *Zenodo.* https://doi.org/10.5281/zenodo.6831287.

Panda, R., Sethi, M., and Agrawal, S. (2018). Sustainable Development Goals and India: A Cross-Sectional Analysis. *OIDA International Journal of Sustainable Development, 11*(11), 79-90. Retrieved from https://ssrn.com/abstract=3308074

Sharma, D., and Jain, S. (2015). Evaluation of Stemming and Stop Word Techniques on Text Classification Problem. *International Journal of Scientific Research in Computer Science and Engineering, 3*(2), Page Range. ISSN: 2320-7639. Retrieved from www.isroset.org.

Smith, T. B., Vacca, R., Mantegazza, L., *et al.* (2021). Natural Language Processing and Network Analysis provide Novel Insights on Policy and Scientific Discourse around Sustainable Development Goals. *Scientific Reports, 11*(1), 22427. https://doi.org/10.1038/s41598-021-01801-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *In Proceedings of the Conference on Neural Information Processing Systems.*

Vayansky, I., and Kumar, S. A. P. (2020). A Review of Topic Modeling Methods. *Information Systems, 94, 101582.*

Webster, J., and Kit, C. (1992). Tokenization as the Initial Phase in NLP. *Proceedings of the 14th Conference on Computational linguistics - Volume 4,* 1106-1110. https://doi.org/10.3115/992424.992434.

Zhou, X., Jain, K., Moinuddin, M., and McSharry, P. (2022). Using Natural Language Processing for Automating the Identification of Climate Action Interlinkages within the Sustainable Development Goals. *In Association for the Advancement of Artificial Intelligence (AAAI) 2022 Fall Symposium on the Role of AI in Responding to Climate Challenges.*

Annex

Table A1: Description of SDGs

SDG	Description
1. No Poverty	End poverty in all its forms everywhere.
2. Zero Hunger	End hunger, achieve food security, and promote sustainable agriculture.
3. Good Health and Well-being	Ensure healthy lives and promote well-being for all at all ages.
4. Quality Education	Ensure inclusive and equitable quality education and promote lifelong learning opportunities.
5. Gender Equality	Achieve gender equality and empower all women and girls.
6. Clean Water and Sanitation	Ensure availability and sustainable management of water and sanitation for all.
7. Affordable and Clean Energy	Ensure access to affordable, reliable, sustainable, and modern energy for all.
8. Decent Work and Economic Growth	Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all.
9. Industry, Innovation, and Infrastructure	Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.
10. Reduced Inequality	Reduce inequality within and among countries.
11. Sustainable Cities and Communities	Make cities and human settlements inclusive, safe, resilient, and sustainable.
12. Responsible Consumption and Production	Ensure sustainable consumption and production patterns.
13. Climate Action	Take urgent action to combat climate change and its impacts.
14. Life Below Water	Conserve and sustainably use the oceans, seas, and marine resources for sustainable development.
15. Life on Land	Protect, restore, and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification and halt and reverse land degradation and halt biodiversity loss.
16. Peace, Justice, and Strong Institutions	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels.
17. Partnerships for the Goals	Strengthen the means of implementation and revitalize the global partnership for sustainable development.

Source: https://sdgs.un.org/goals.

I. Primer on NLP

a. Terminology

The fundamental unit of analysis in NLP is the "term", which can be either a single word like "development" or a sequence of words known as "n-grams" (Cavnar *et al.,* 1994). While words retain their conventional meaning, n-grams represent groups of words treated as a single unit for analysis. Examples include "Good Health" and "Quality Education" which are "bi-grams" (Hachaj *et al.,* 2018) often analysed as single terms within NLP tasks.

Documents can range from sentences and paragraphs to entire literary works and are typically the primary unit of analysis when analysing text. Here, the documents which are analysed are budget speeches from the Union Government and six States for fiscal years 2012 to 2023.

Finally, a corpus is a collection of documents analysed, equivalent to a 'data set'. In this case, the corpus is the set of budget speeches. The corpus is comprised of multiple documents, each of which is comprised of terms. Anastasopoulos *et al.,* (2017) neatly summarised the NLP processing hierarchy in mathematical notations (Chart A1).

Chart A1: NLP Processing Hierarchy

Terms ⊆ Document ⊆ Corpus

Source: Anastasopoulos *et al.*, (2017); and Authors' illustration.

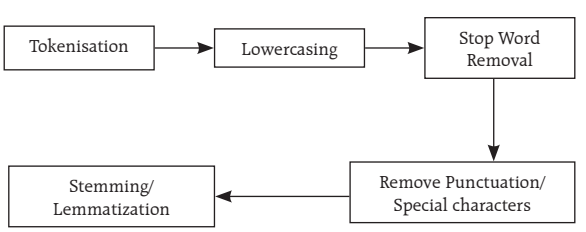
b. From Text to Data

Text pre-processing is a critical step in NLP (Kurniasih *et al.*, 2022). It involves several key steps to refine the raw text data before it is converted to numbers. The initial step is tokenisation where text is segmented into individual words or tokens (Webster *et al.*, 1992). Lowercasing is essential for ensuring text consistency by converting all characters to lowercase, making "budget" and "Budget" equivalent in analysis. Stop word removal (Sharma *et al.*, 2015) filters out common, low-information words like "the", "is", "are", *etc.*, so that the focus is on the words that convey meaning. Removing punctuation, such as periods or commas, improves text readability and removes unnecessary clutter. Lastly, stemming or lemmatization (Khyani *et al.*, 2021) simplifies words to their root forms, making sure that words with similar meanings are represented in the same way; "budgeting", "budgeted" and "budgets" all become "budget" (Chart A2).

c. Document–Term Matrix (DTM)

A DTM is a key tool in NLP and text analysis, transforming documents into a structured numerical format for analysis (Antonellis *et al.*, 2006; and Anastasopoulos *et al.*, 2017). It represents each document in rows and unique terms in columns, using values to indicate term frequency or presence in documents. Different methods can be used to fill the cells, including TF (term frequency), binary (presence/absence), and TF-IDF (combining term frequency with term uniqueness) [Christian *et al.*, 2016]. The most common of these is TF-IDF which will be discussed in detail in the next section.

Chart A2: Text Pre-Processing Pipeline



Sources: Camacho-Collados, J. *et al.*, (2017); and Authors' illustration.

d. Overview of TF-IDF

The key intuition behind TF-IDF is that a term's importance is inversely related to its frequency across documents. This score combines two factors: Term Frequency (TF) measures how often a term appears in a document, while Inverse Document Frequency (IDF) penalizes common terms across the entire corpus. IDF is calculated as:

$$IDF_{t,D} = log\left(\frac{N}{count(d \epsilon D : t \epsilon d)}\right) \qquad \text{Equation (1)}$$

In this context, *t* represents the term (word) for which this paper aim to assess commonality, and *N* stands for the total number of documents (d) in the corpus (D). The denominator corresponds to the count of documents where the term *t* is present.

To prevent divide-by-zero errors when terms are absent in the corpus, IDF calculations typically add 1 to the count of documents containing the term, effectively adjusting the denominator to (1 + count). The popular library scikit-learn solves this by modifying the formula as follows:

$$IDF_t = 1 + log\frac{(1+n)}{(1+DF_t)} \qquad \textit{Equation (2)}$$

Thus, with IDF, infrequent terms stand out, revealing hidden insights within documents.

Putting it together: TF-IDF

TF-IDF scores, calculated by multiplying TF and IDF.

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D} \qquad \text{Equation (3)}$$

A term's relevance within a document is reflected by its TF-IDF score, with 0 representing minimal importance and higher scores signifying increasing significance. This article uses c-TF-IDF which is a class-based TF-IDF procedure that can be used to generate features from textual documents based on the class they are in[7].

[7] https://github.com/MaartenGr/cTFIDF