Trilingual Topic Modeling of Sri Lankan Parliamentary Debates

MERCon 2026

HImath Dhanapala · Haren Daishika · Himandhi Kuruppu · Sithija Seneviratne · Ashini Kavindya · Patalee Narasinghe · Sandeepa Weerasekara · Nisansa de Silva · Sandareka Wickramanayake

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING, UNIVERSITY OF MORATUWA

Presented by Himath Dhanapala & Himandhi Kuruppu

The Problem

1. Trilingual and code-mixed discourse makes language-independent topic discovery difficult.

2. Sinhala and Tamil morphology fragments vocabulary and weakens bag-of-words topic models.

3. Dual-column, mixed-script Hansard PDFs make reliable text extraction difficult for conventional OCR

Research Questions

RQ1

Can one pipeline model a trilingual, code-mixed corpus while preserving cross-lingual meaning?

RQ2

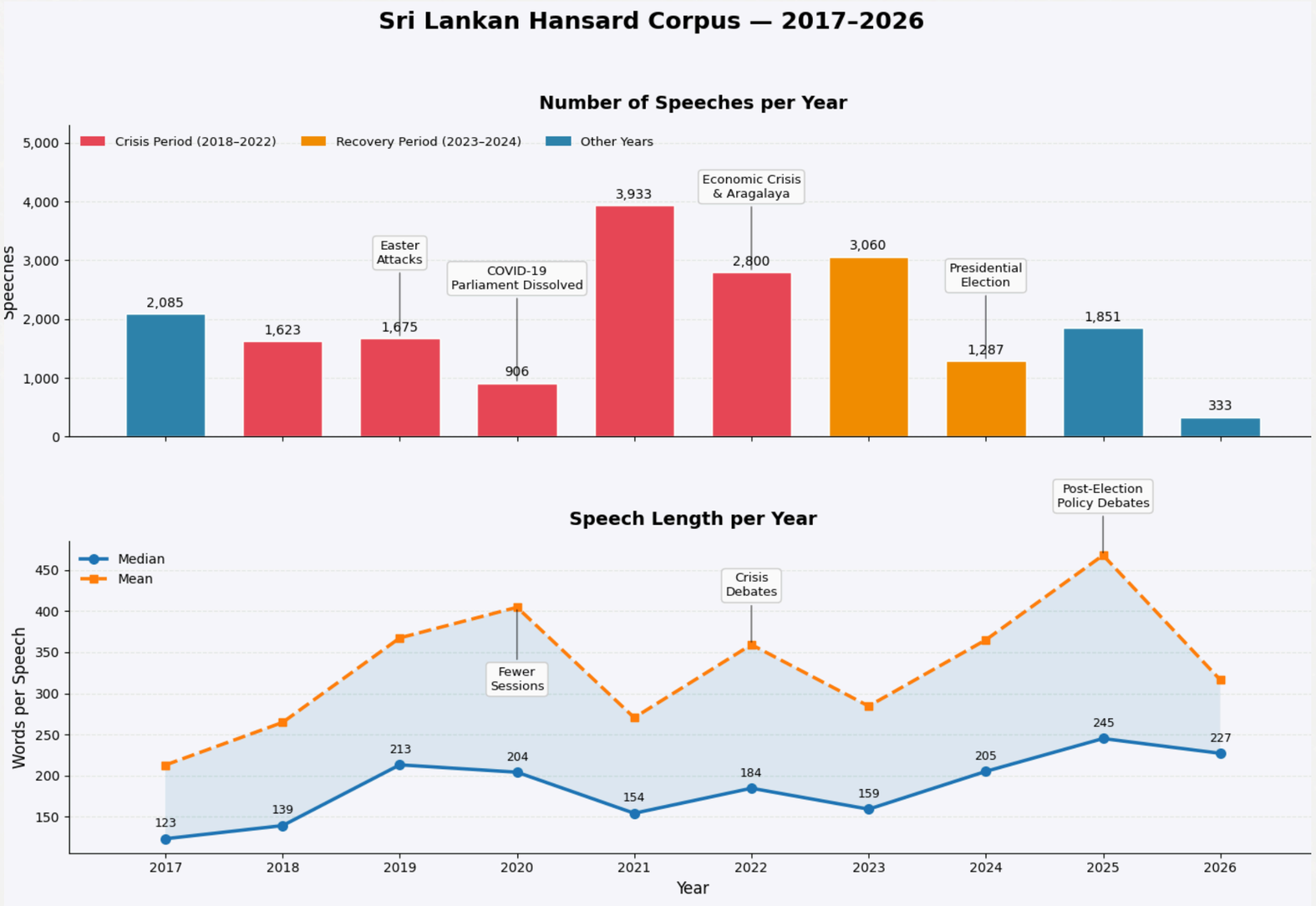
Can macro-topics be derived empirically from micro-topic structure, rather than by choosing K?

RQ3

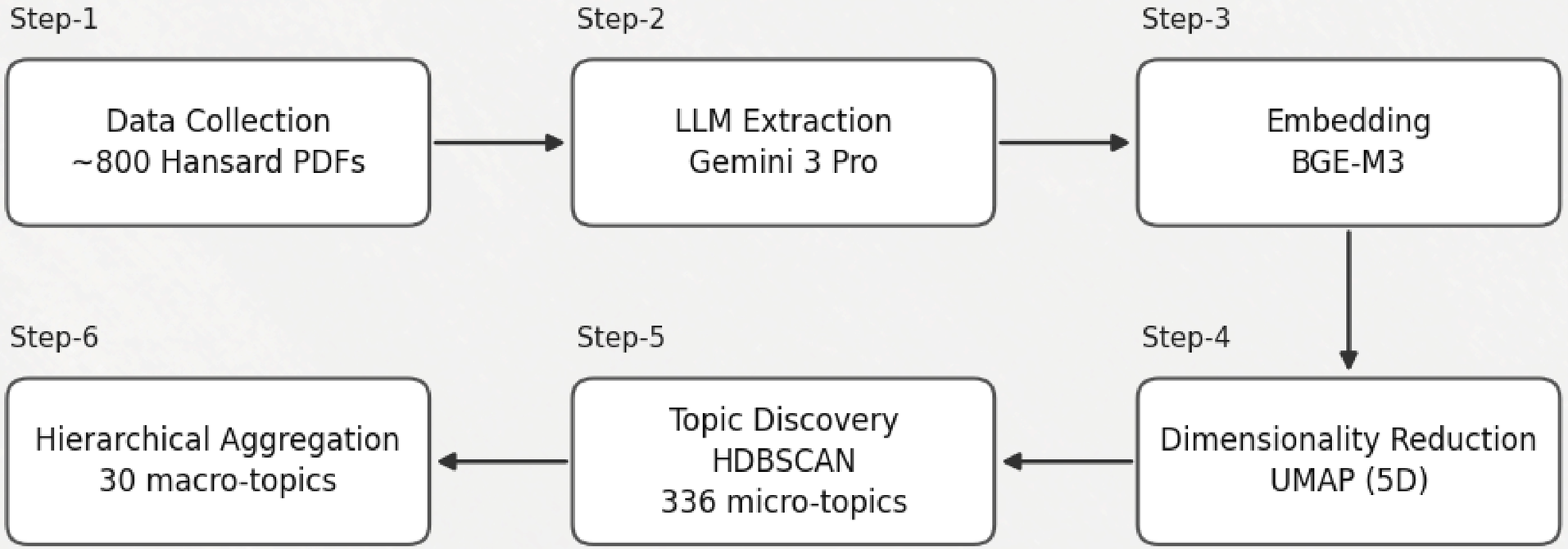
Does adding lexical information to dense embeddings improve topic separation?

Corpus & Extraction

- ~800 Hansard PDFs scraped from the official parliamentary archive, 2017–2026
- Gemini 3 Pro used for multilingual spee extraction
- 19,553 substantive speeches 76% Sinhala, 7% Tamil, 7% English, 10% code-mixed
- 1,045-term trilingual stopword list built for parliamentary boilerplate



The Pipeline



Choosing the Embedding Model

High similarity alone is not enough

E5 scores highest on STS, but its anisotropy hurts clustering and its 512-token limit truncates 35-40% of speeches while BGE-M3 supports 8,192 tokens.

Model	STS	MRR	P@5	Anisotropy
multilingual-e5-large	0.827	0.810	0.56	0.909
BAAI/bge-m3	**0.523**	**0.516**	**0.68**	**0.552**
LaBSE	0.487	0.324	0.72	0.492
mpnet-base-v2	0.424	0.536	0.76	0.496

Choosing the Clustering method

Higher purity comes at a coverage cost.

HDBSCAN achieves the highest cluster purity and separation, while rejecting 41.7% of weakly topical speeches as noise.

Algorithm	K	Noise %	B-Cubed P	Silhouette
K-Means	13	0.0	0.349	0.388
Agglomerative (Ward)	13	0.0	0.355	0.377
HDBSCAN	**11**	**41.7**	**0.673**	**0.532**
OPTICS	19	36.1	0.670	0.497

From 336 micro topics to 30 macro topics

Step 1

HDBSCAN yields 336 micro-topics

Step 2

Pareto filtering retains the dominant micro-topics

Step 3

Cluster their centroids using cosine distance + average linkage

Step 4

Largest merge-distance gap gives 30 macro-topics

Are the Topics Interpretable?

Word clouds

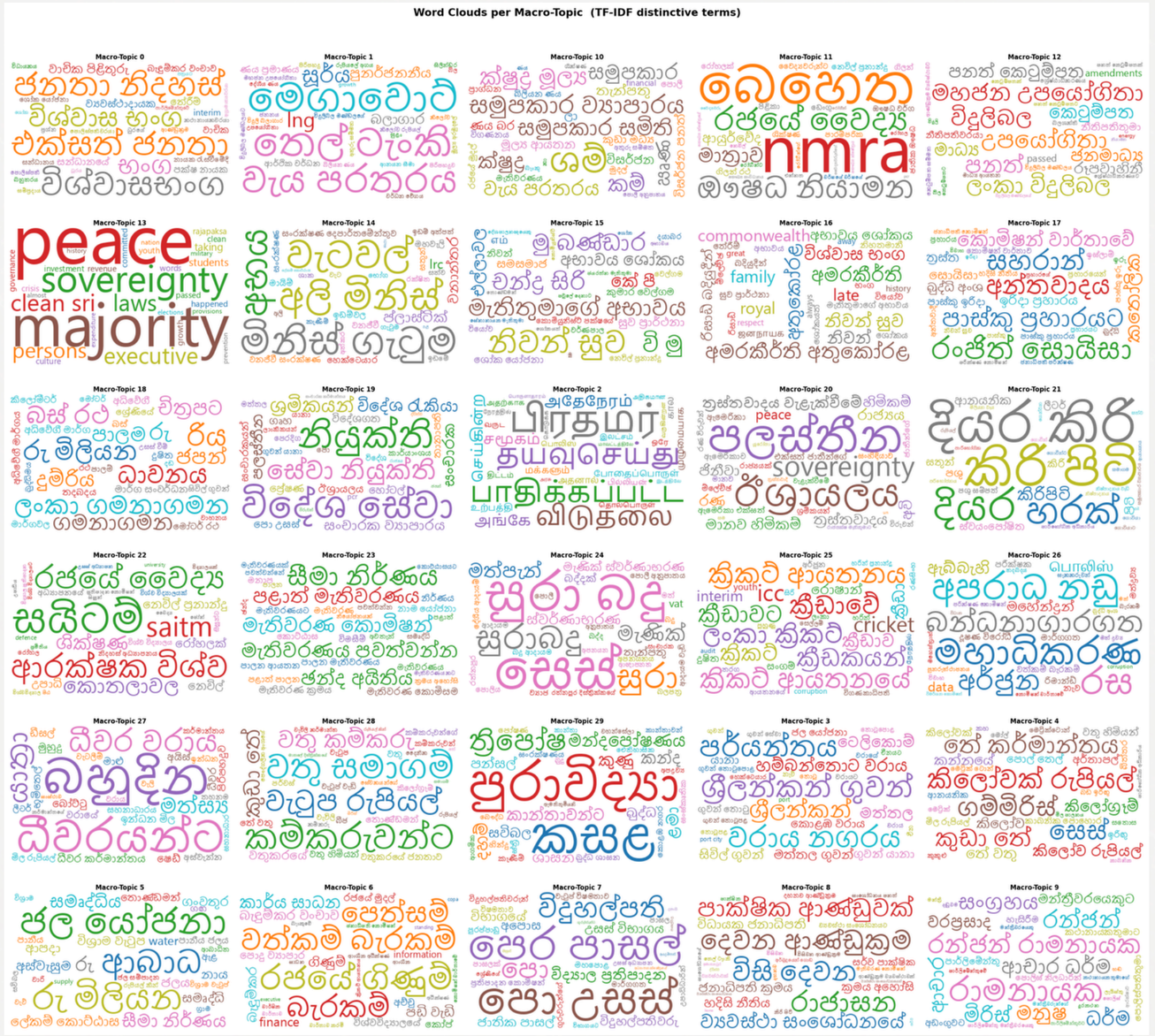
30 macro-topics, distinctive terms

Cross-lingual

Most topics span multiple languages

Externally plausible

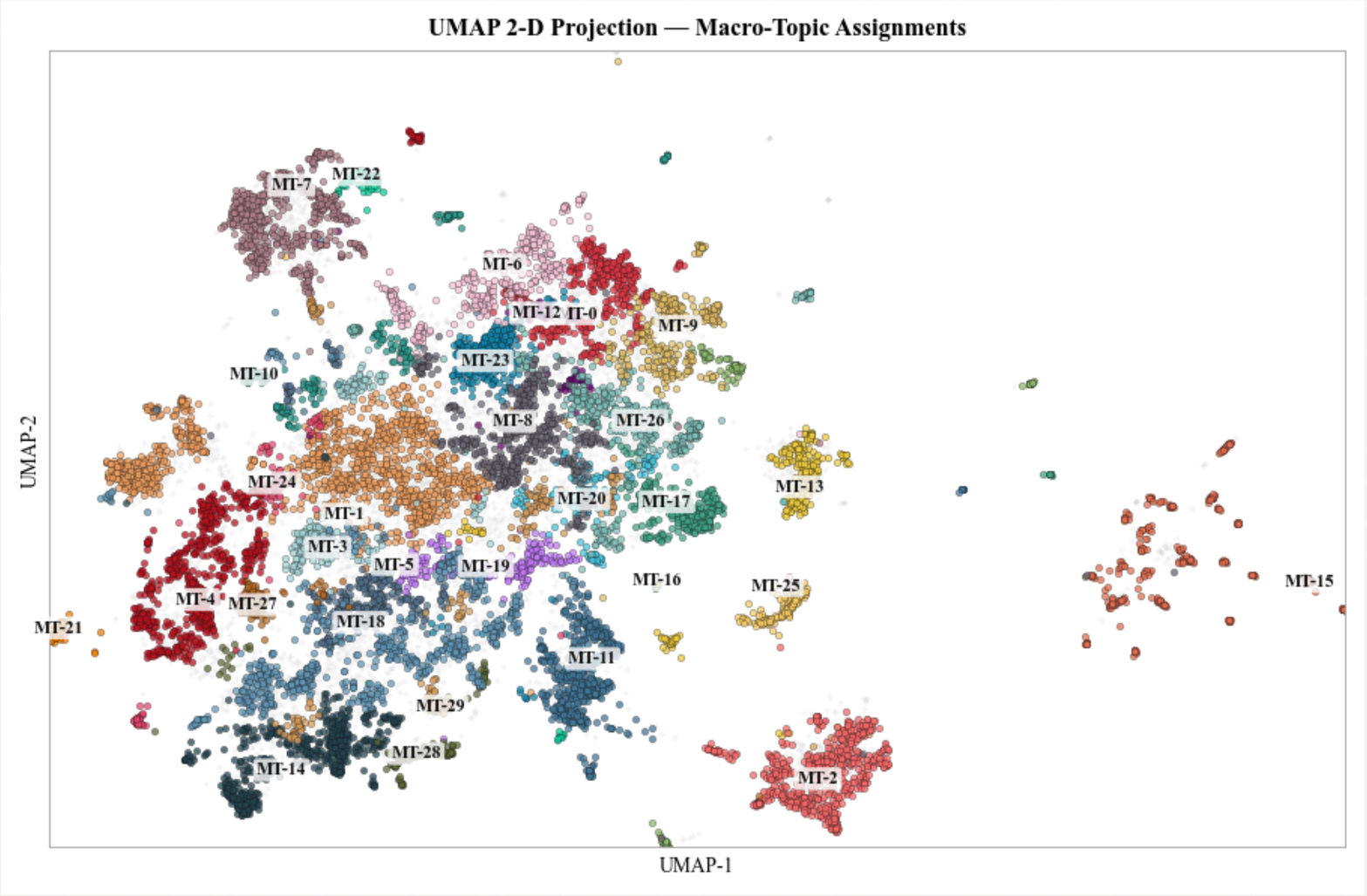
Temporal patterns align with real events



Do the Topics Hold Up?

UMAP projection

Many macro-topics occupy distinct regions

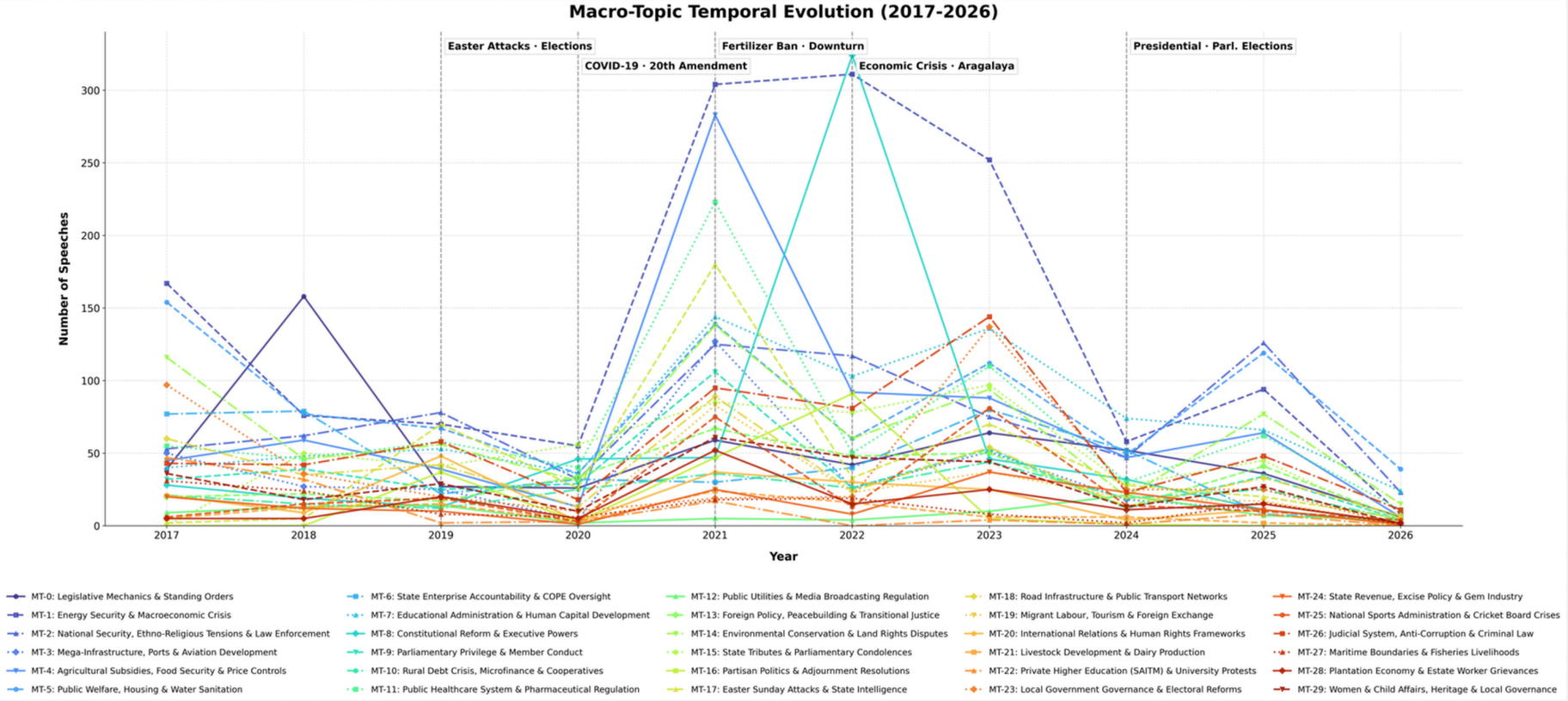


Temporal Validation

No event labels were used.

- Economy peaks during the 2022 crisis / Aragalaya
- Easter Sunday discourse rises around 2019
- Healthcare peaks during COVID-19

The temporal alignment emerges from the discourse itself.



Findings, Limitations & Next Steps

30 data-driven macro-topics reveal cross-lingual, event-aligned parliamentary themes without supervision.

Key Findings

- 336 micro-topics to 30 macro-topics
- Most topics span multiple languages
- Temporal peaks align with major national events
- BiTopic consolidates topics, but increases noise.

Limitations

- Gold set: 300 speeches, 91% Sinhala
- Rare categories were excluded from evaluation
-

Future Work

- Aspect-based sentiment within topics
- Temporal sentiment across sessions
- Further refinement of BiTopic fusion

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[2] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.

[3] D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, "Word embedding evaluation for Sinhala," in Proc. 12th Language Resources and Evaluation Conf. (LREC), 2020, pp. 1874-1881.

[4] Y. Wijeratne and N. de Silva, "Sinhala language corpora and stopwords from a decade of Sri Lankan Facebook," arXiv preprint arXiv:2007.07884, 2020.

[5] F. R. Baumgartner, C. Breunig, and E. Grossman, Comparative Policy Agendas: Theory, Tools, Data. Oxford, U.K.: Oxford Univ. Press, 2019.

[6] N. de Silva, "Survey on publicly available Sinhala natural language processing tools and research," arXiv preprint arXiv:1906.02358v26, 2026.

[7] R. Smith, "An overview of the Tesseract OCR engine," in Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR), vol. 2, 2007, pp. 629-633.

[8] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Inf. Sci., vol. 622, pp. 178-210, 2023.

References

[9] Y. Li and H. Wu, "A clustering method based on K-means algorithm," Phys.Procedia, vol. 25, pp. 1104-1109, 2012.
[10] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density basedclustering," J. Open Source Softw., vol. 2, no. 11, p. 205, 2017.
[11] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering basedon hierarchical density estimates," in Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), 2013, pp. 160-172.
[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982-3992.
[13] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 16857-16867.
[14] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2022, pp. 878-891.
[15] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," arXiv preprint arXiv:2402.03216, 2024.
[16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," arXiv preprint arXiv:2402.05672, 2024.
[17] A. F. AlShammari, "Implementation of keyword extraction using term frequency-inverse document frequency (TF-IDF) in Python," Int. J. Comput. Appl., vol. 185, no. 35, pp. 9-14, 2023.

References

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
[19] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," in Proc. 16th Conf. European Chapter Assoc. Comput. Linguistics (EACL), 2021, pp. 1676-1683.
[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993-1022, Jan. 2003.
[21] N. I. Senaratna, "Sri Lanka document datasets: A large-scale, multilingual resource for law, news, and policy," arXiv preprint arXiv:2510.04124, 2025.
[22] B. E. Lauderdale and T. S. Clark, "Scaling politically meaningful dimensions using texts and votes," Amer. J. Political Sci., vol. 58, no. 3, pp. 754-771, 2014.
[23] G. Abercrombie and R. T. Batista-Navarro, "A sentiment-labelled corpus of Hansard parliamentary debate speeches," in Proc. LREC 2018 Workshop ParlaCLARIN, 2018, pp. 43-47.
[24] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in Proc. Machine Translation Summit X, 2005, pp. 79-86.
[25] B. H. R. Pushpananda, "Improving performance of statistical machine translation between morphologically rich and low resourced language pairs," Ph.D. dissertation, Univ. of Colombo School of Computing, Colombo, Sri Lanka, 2017.
[26] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., "Gemini: A family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.
[27] Parliament of Sri Lanka, "Hansard: Official reports of parliamentary debates," 2026. [Online]. Available: https://www.parliament.lk/en/business-of-parliament/hansards

Thank You!