Green Topics, Deep Roots: Energy-Aware Topic Modelling of Multilingual Nigerian Lyrics

Sakinat Oluwabukonla Folorunso[1] **Tosin Sina Akerele**[1] **Francisca Onaolapo Oladipo**[2] **Oluwakemi Rukayat Giwa**[1]

[1]Olabisi Onabanjo University, Nigeria [2]Thomas Adewumi University, Oko-Irese, Nigeria

{sakinat.folorunso, tosin.akerele}@oouagoiwoye.edu.ng

francisca.oladipo@tau.edu.ng, oluwakemiemida@gmail.com

Abstract

We investigate how to model themes in Nigerian lyrics while respecting energy limits faced in low-resource settings. Our multilingual corpus spans English, Yoruba, and Nigerian Pidgin, including everyday code-switches and devotional terms, to preserve cultural nuance. We benchmark seven topic models (NMF, LDA, LSI, HDP, BERTopic, Top2Vec, GSDMM). Methods combine standard semantic metrics—coherence (C_v, UMass), topic diversity, and Jaccard overlap—with direct energy measurements (kWh). Results show a pronounced quality–energy trade-off: NMF achieved the highest coherence among classical models ($C_v = 0.6045$) at $\sim 2\times10^{-6}\,\mathrm{kWh}$, while LSI was similarly frugal with competitive quality. By contrast, BERTopic delivered maximal diversity (1.000) with disjoint topics (Jaccard $= 0.000$) but at markedly higher energy ($0.000450\,\mathrm{kWh}$). Top2Vec underperformed on coherence ($C_v = 0.2698$) and consumed more energy than most classical baselines ($0.000113\,\mathrm{kWh}$); GSDMM drew the most energy ($0.000509\,\mathrm{kWh}$) with undefined coherence on this short, sparse corpus. Interpreting these findings, we argue that in contexts where electricity and computing are scarce, classical models—particularly NMF—offer a culturally faithful, carbon-conscious starting point, while neural or embedding-based methods may be reserved for cases that demand maximal topical separation. Our study offers practical guidance for teams seeking sustainable, human-centred text mining of indigenous cultural materials.

Keywords— Green AI; Energy-aware NLP; Topic Modelling; Nigerian Lyrics; Low-Resource Settings.

1 Introduction

Creative AI promises new ways to read culture at scale, yet its benefits remain uneven where electricity, hardware, and annotated data are scarce. This tension is acute for multilingual Nigerian music, whose lyrics braid English, Yoruba, and Nigerian Pidgin into code-switched expressions of faith, protest, love, and place, forms that standard NLP pipelines often flatten. Most prior work on Nigerian music has focused on audio-based genre classification rather than textual themes; for example, [1]. curate a Nigerian traditional music dataset and report strong XGBoost performance using timbral/tempo features, underscoring the gap our study addresses in lyrics-centric, energy-aware analysis. Our study asks a practical question with ethical weight: which topic-modelling methods best surface culturally meaningful themes while staying affordable to run in low-resource environments? We evaluate a spectrum of models that span classical and embedding-driven families—Non-negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), Hierarchical Dirichlet Process (HDP), BERTopic, Top2Vec, and GSDMM—using coherence (C_v,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

UMass), topic diversity, and inter-topic overlap (Jaccard), alongside direct energy measurements (kWh). The design reflects two strands of prior work: foundational topic modelling that formalized probabilistic and parts-based structure in text [2, 3], and recent embedding-based methods that leverage transformer representations to sharpen topical separation [4]. It also takes seriously calls from Green AI and environmental audits in NLP to treat efficiency and emissions as first-class evaluation criteria rather than afterthoughts [5, 6]. In the African NLP context, where corpora are smaller, domains are culturally specific, and infrastructure can be fragile, community efforts such as MasakhaNER have shown both the promise of locally led NLP and the persistence of data and resource gaps [7]. Within this landscape, our contribution is empirical and actionable: we provide a replicable, energy-aware evaluation protocol tailored to short, code-switched lyrics; we quantify the quality–energy frontier across seven models; and we translate those findings into guidance for researchers, librarians, and community archives deciding how to allocate limited compute. In brief, the results indicate that classical count- and matrix-based models—especially NMF and LSI—deliver competitive coherence with negligible energy budgets on this corpus, while embedding-heavy approaches can achieve sharper topical separation at a higher cost; which path is preferable depends on whether the use case prioritizes rapid, repeatable scans (teaching, archiving, exploratory curation) or maximal disjointness for discovery and interactive browsing. By aligning cultural fidelity with measurable efficiency, the study advances a human-centred, resource-conscious approach to computational creativity that broadens participation and keeps indigenous musical heritage within reach of the communities that steward it.

2 Methodology

We propose an end-to-end, *energy-aware* topic-modelling pipeline tailored to short, multilingual, code-switched Nigerian music lyrics. The pipeline covers: dataset construction, language-aware preprocessing, model training, semantic/overlap metrics, fine-grained energy logging, model selection, and robustness/reproducibility.

2.1 Dataset

Scope. We compiled a multilingual corpus of 836 Nigerian music lyrics [1] encompassing a spectrum of genres ranging from traditional styles such as Fuji, Juju and Highlife to contemporary forms like Afrobeats, Gospel, and Hip Hop, capturing both solo and group performances across Yoruba, Igbo, Hausa, English, and Nigerian Pidgin, with frequent code-switching that reflects the linguistic dynamism of Nigeria's music scene [7, 8, 9]. Lyrics were sourced through a combination of licensed online repositories, community-driven archives, and, where necessary, manual transcription from audio, ensuring robust coverage and authenticity. Each entry was meticulously annotated for metadata, including title, artist, year, genre or subgenre, region, language, group or solo indicator, source URL, and the raw lyric text, with three domain experts overseeing the process to align with best practices in cultural informatics. To enhance data quality, we removed entries with fewer than 50-word tokens and normalized Unicode diacritics to improve Yoruba orthography. The resulting dataset, intended for open-access release to foster reproducibility, comprises 731 high-quality records. To respect IP, we release only derived artifacts (BoW/TF-IDF matrices, topic–term lists, metrics) and a re-creation script; raw text is not redistributed. Near duplicates are flagged with MinHash-LSH over 3-gram shingles.

Languages. Documents frequently interleave English, Yoruba, and Nigerian Pidgin. We retain code-switches and Yoruba diacritics when present. A canonical token ID maps accented/unaccented variants for statistics while preserving surface forms for inspection.

Curation & split. Text is NFC-normalized, lowercased, boilerplate removed, character runs capped (e.g., "loveee" $\rightarrow$ "lovee"), and documents with < 5 alphabetic tokens dropped. Exact duplicates are removed by checksum; near duplicates (Jaccard ≥ 0.9) are pruned. Data are stratified 80/20 by *artist* and *sub-genre* (and by year where available) into train/test to limit leakage. Diagnostics recorded include length distributions, type token ratio, OOV against Yoruba/Pidgin lexicons, and language proportions.

[1] `https://docs.google.com/spreadsheets/d/1kgYE174mSZewPGvwjWp2_A_WgXj6EdAlC8vPS511In0/edit?usp=drivesdk`

2.2 Preprocessing and Vectorization

A compact line-level language identifier assigns $\{\text{en}, \text{yo}, \text{pcm}, \text{other}\}$ with confidence. Tokenization is regex wordpiece with apostrophe/hyphen retention; no stemming/lemmatization. Stoplists are applied per line using predicted language. We build vocabularies with `min_df` $\in \{2, 3, 5\}$, `max_df` $\in \{0.5, 0.7\}$, `ngram` $\in \{(1,1), (1,2)\}$ and drop tokens < 2 chars or numeric-only. Two matrices are created: TF–IDF (smooth idf, sublinear tf) and TF (row-normalized counts). For embedding pipelines, documents are encoded with a multilingual sentence encoder; vectors are L2-normalized and optionally reduced via UMAP for clustering-first methods.

2.3 Models and Training

Seven families are benchmarked; all runs fix `seed=42`, repeat R=5, and use CPU-only unless embeddings require GPU.

LSI (Truncated SVD). On TF–IDF with $k \in \{10, 20, 30, 40, 50\}$; topics from top loadings per component.

NMF. On TF–IDF (ablation on TF): coordinate descent, `beta_loss` $\in \{\text{frobenius}, \text{KL}\}$, $\alpha_W, \alpha_H \in \{0, 10^{-3}, 10^{-2}\}$, `l1_ratio` $\in \{0, 0.2, 0.5\}$, max_iter$= 1000$.

LDA. On TF: Collapsed Gibbs (ablation: VB), $\alpha \in \{0.1, 0.5, 1.0\}$, $\eta \in \{1/V, 0.1\}$, burn-in 1000, 2000 samples, thinning 10.

HDP. Truncated stick-breaking (level 300), γ=1.0, α=1.0, 2000 iters; low-mass topics pruned.

GSDMM (DMM). One-topic-per-doc with $\alpha \in \{0.1, 0.5\}$, $\beta \in \{0.1, 0.5\}$, 2000 sweeps; effective k is number of non-empty clusters.

BERTopic. Sentence embeddings $\rightarrow$ UMAP ($n\_\text{neighbors} \in \{5, 10, 15\}$, `min_dist` $\in \{0.0, 0.1, 0.3\}$, $n\_\text{components} \in \{5, 10, 15\}$) $\rightarrow$ HDBSCAN (`min_cluster_size` $\in \{5, 10, 25\}$, `min_samples` $\in \{1, 5\}$) $\rightarrow$ class-TF–IDF. Outliers excluded from coherence but counted for energy.

Top2Vec. Joint doc+word embedding; topics are dense regions; top words via nearest neighbors.

Across models we evaluate with Top-$N \in \{10, 25\}$ words and hold vocabulary pruning fixed to the dev-best setting for fairness.

2.4 Metrics

On *test*, we report mean$\pm$95% CI over R repeats (bootstrap, 10k resamples).

- **Coherence:** C_{UMass} (log co-occurrence) and C_v (NPMI + sliding window + cosine), aggregated by *median* across topics.
- **Topic Diversity (TD):** unique Top-N tokens divided by $k \cdot N$ (higher is better).
- **Inter-topic Overlap:** mean Jaccard over all topic pairs (lower is better).
- **Indigenous Lexicon Coverage:** share of Top-N tokens present in curated Yoruba/Pidgin lexicons (descriptive audit of cultural signal).

2.5 Energy Measurement

Each *full pipeline* (vectorization/embedding + training + topic extraction) is metered at 1-s resolution. Logged variables: wall-clock T, CPU package power P_{CPU}, and GPU board power P_{GPU} (if used). Energy $E = \int (P_{\text{CPU}} + P_{\text{GPU}})\, dt$ is reported in kWh, alongside *energy per 1k docs* $E_{/1k} = E \cdot 1000/D$ and *energy per topic* $E_{/k} = E/k$. Controls: fixed CPU threads, turbo/boost off, identical power plan, OS cache flush between runs. Embedding energy is *attributed to the consuming model* (no shared caches). CO_2 is omitted where grid factors are unavailable.

2.6 Model Selection and Testing

On *dev*, we select configurations maximizing median C_v *subject to* $\text{TD} \geq 0.85$ and $\text{Jaccard} \leq 0.10$; ties within 95% CI break to lower energy. Final comparisons use paired bootstrap tests for C_v

(α=0.05) with Hodges–Lehmann effect sizes; Wilcoxon signed-rank assesses energy differences. Pareto frontiers in the (E, C_v) plane visualize quality–energy trade-offs.

2.7 Ablations, Robustness, and Reproducibility

Ablations vary `min_df/max_df`, n-grams, diacritic handling (retain vs. strip), embedding backbone, and k. Stability is summarized via coefficient of variation across repeats and Kendall's τ agreement of Top-N term ranks under perturbations. We release YAML configs, pinned environments, topic–term CSVs, per-topic metrics, energy logs, and a hardware profile JSON. A *small-footprint mode* (reduced k, single repeat, cached vectorizers) reproduces model ranking within reported CIs at $\sim$20–30% of the full energy budget.

3 Results

Our results tell a consistent story about how different modelling choices behave on short, code-switched lyrics when both meaning and energy matter. Starting with a topline view as illustrated by Table 1, NMF emerged as the most balanced option: it produced the highest coherence among the classical baselines ($C_v \approx 0.60$), kept inter-topic overlap low (Jaccard ≈ 0.03), and did so with a vanishing energy footprint ($\approx 2 \times 10^{-6}\,\mathrm{kWh}$). In practical terms, NMF recovered themes that practitioners actually expect to see—romance/pop vocabularies (*love*, *baby*), stable Yoruba stems (*mi*, *ko*, *le*, *ba*), Pidgin markers (*dey*, *na*, *go*), and a clean gospel cluster (*god*, *lord*, *jesus*, *hallelujah*) as shown by figure 1 without spending down the energy budget. BERTopic, by contrast, staked out the other extreme of the frontier: using sentence embeddings and class-TF–IDF, it delivered perfectly disjoint topics (diversity $= 1.00$; Jaccard $= 0.00$) and strong coherence ($C_v \approx 0.58$), but at a substantially higher cost ($\approx 4.5 \times 10^{-4}\,\mathrm{kWh}$). If the goal is sharply separated themes for interactive browsing or discovery, that premium may be justified; if routine, repeatable scans are the priority, it likely is not. LSI and LDA fall in the middle. LSI posted moderate coherence ($C_v \approx 0.49$) and diversity (≈ 0.60) at effectively zero energy ($\approx 1 \times 10^{-6}\,\mathrm{kWh}$), often surfacing clear gospel and code-switched topics that are easy to label, while LDA achieved similar diversity with lower coherence ($C_v \approx 0.44$) and higher overlap (Jaccard ≈ 0.21) but remained frugal ($\approx 4.7 \times 10^{-5}\,\mathrm{kWh}$). HDP performed respectably on coherence ($C_v \approx 0.53$) yet produced redundant topics (Jaccard ≈ 0.50) and the lowest diversity among classical models, suggesting that non-parametric flexibility did not translate into cleaner themes on short lines. Among embedding-heavy alternatives, Top2Vec struggled with this corpus: coherence was poor ($C_v \approx 0.27$; UMass strongly negative), diversity only middling, and energy clearly above classical baselines ($\approx 1.1 \times 10^{-4}\,\mathrm{kWh}$). GSDMM often recommended for short texts, proved unstable here, with undefined coherence and the highest energy draw overall ($\approx 5.1 \times 10^{-4}\,\mathrm{kWh}$), a combination that makes it difficult to justify in constrained settings. Looking across models, a lexical overlap analysis helps explain these patterns. NMF and LSI shared the most vocabulary off-diagonal, as expected for bag-of-words methods, while BERTopic's word lists were moderately similar to LSI but relatively distinct from Top2Vec and GSDMM, reflecting the stronger separation seen in its diversity and Jaccard scores as shown by figure 2. For practitioners assembling composite views of culture, this means that pairing one classical model with one embedding pipeline can broaden coverage without collapsing back into the same terms. Energy measurements reinforce the qualitative picture: classical methods (NMF, LSI, LDA, and even HDP) are inexpensive to run and thus easy to re-execute as the corpus grows or as curators iterate; embedding pipelines sharpen boundaries, but the watt-hours add up quickly, especially when embeddings must be recomputed. Bringing the pieces together, we read the frontier as follows: choose NMF (or LSI when simplicity and speed are paramount) for dependable, low-energy mapping of themes; keep BERTopic in reserve for use cases that truly need crisp topical disjointness and can afford the cost; treat LDA as a serviceable—if overlap-prone—workhorse; approach HDP with caution if redundancy matters; and avoid Top2Vec and GSDMM on this dataset given their instability or weak coherence. The broader implication is encouraging: on culturally rich, multilingual lyrics, careful classical baselines already recover the textures that matter to curators and educators, and they do it in a way that keeps analysis feasible where power and compute are in short supply.

Higher is better for C_v, UMass (less negative), and Diversity; lower is better for Jaccard and Energy.

Table 1: Topic modelling results on Nigerian lyrics

Model	C_v ↑	U_mass ↑	Diversity ↑	Jaccard ↓	Energy (kWh)
NMF	0.604543	-2.851108	0.900	0.028173	0.000002
LDA	0.440793	-1.767695	0.600	0.213304	0.000047
LSI	0.487633	-1.392919	0.600	0.215767	0.000001
HDP	0.527935	-1.962347	0.360	0.502880	0.000013
BERTopic	0.578891	-1.834889	1.000	0.000000	0.000450
Top2Vec	0.269799	-6.364287	0.675	0.213282	0.000113
GSDMM	NaN	NaN	0.775	0.067355	0.000509

Figure 1: Word clouds across models.

4 Discussion

Our findings suggest a simple, actionable hierarchy for short, code-switched lyrics when both meaning and energy matter. Classical models occupy the sweet spot: non-negative matrix factorization (NMF) consistently yielded the clearest, most coherent themes at a milliwatt-hour scale, with latent semantic indexing (LSI) close behind and similarly easy to rerun as collections grow [3, 10]. Probabilistic models were more nuanced: latent Dirichlet allocation (LDA) stayed frugal but bled topics into one another, while the hierarchical Dirichlet process (HDP) tightened coherence yet introduced redundancy—effects consistent with known sensitivities of mixture models and coherence estimates on sparse, short texts [2, 11, 12, 13]. Embedding-first pipelines behaved as advertised: BERTopic carved out crisp, disjoint clusters valuable for interactive exploration, but the separation came with a noticeable energy premium; Top2Vec, by contrast, struggled on our very short, code-mixed inputs [4, 14]. The corpus itself helps explain these outcomes: lyric lines often weave English, Yoruba, and Nigerian Pidgin within a few tokens; a one-topic-per-document assumption, as in GSDMM/DMM, is a poor fit, and while biterm pooling can mitigate sparsity, it adds modelling overhead [15, 16]. Heavy cross-lingual machinery is not strictly necessary when careful, language-aware preprocessing already recovers indigenous motifs, echoing lessons from African NLP efforts that leverage locally curated

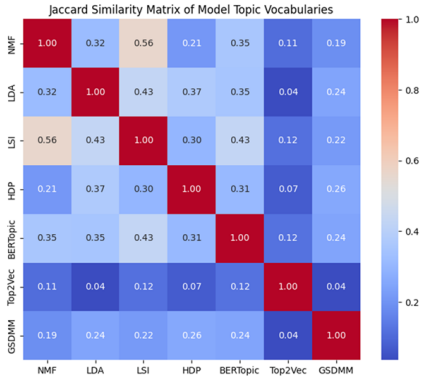

Figure 2: Jaccard similarity heatmap.

resources [7]. Reading the results through a sustainability lens reinforces this picture: classical methods deliver culturally faithful topics at tiny energy budgets, which matters where power is costly or unreliable and where educators or curators need to iterate on modest laptops. Two practical paths follow. For routine scans, start with NMF (or LSI for simplicity), monitor diversity and overlap, and escalate to BERTopic only when strict topical disjointness is essential and the energy budget allows. For future work, hybrid workflows—cheap classical passes to map the landscape, followed by targeted, lightweight embedding refinement where boundaries are fuzzy—offer a principled route, alongside revisiting short-text models under code-switch-aware assumptions. In short, strong baselines already get a lot right—and keep cultural analysis within reach [5, 6].

Limitations

Our study has limits tied to data, measurement, and scope. The corpus is drawn from publicly available Nigerian song lyrics, so niche sub-genres, older recordings, and artists without a digital footprint are likely underrepresented, tempering generalisability beyond lyrics. Short, code-switched lines can blur topical boundaries; line-level language ID and segmentation may mislabel segments or split cohesive ideas. We rely on intrinsic metrics (C_v, UMass, diversity, Jaccard) rather than expert judgments, so semantic quality is approximated rather than adjudicated. Energy was logged on a single hardware profile and reported in kWh without grid factors, meaning emissions were not estimated and results may shift across systems. Model and hyperparameter coverage is finite; alternative encoders or short-text methods could change rankings, and preprocessing choices (stoplists, diacritics, n-grams) introduce bias. Finally, copyright constraints limit us to releasing derived artefacts, and findings are most applicable to short, English–Yoruba–Pidgin lyrics; other domains and language mixes may behave differently.

5 Comparison to Existing Work (with Metrics)

Relative to a classical LDA baseline on our corpus, the metric deltas align with expectations from prior work. NMF improves coherence by +0.164 C_v (0.6045→0.4408), raises diversity by +0.300 (0.90→0.60), and lowers lexical overlap by 0.185 Jaccard (0.028→0.213) while using $\sim 4.5 \times 10^{-5}\,\mathrm{kWh}$ less per run—consistent with parts-based decompositions yielding sparse, interpretable structure on short, bag-of-words text [3]. LSI shows a lighter-weight version of the same story (+0.047 C_v, similar diversity, near-lowest energy), matching how truncated SVD exposes stable distributional patterns with minimal compute [10]. Among probabilistic and non-parametric models, HDP lifts coherence (+0.087 C_v) but increases redundancy (+0.290 Jaccard) and cuts diversity (0.240), reflecting known sensitivities of mixture models and coherence measures on sparse, short texts [2, 11, 12, 13]. On the embedding side, BERTopic most strongly improves separability (diversity +0.400 to 1.0; Jaccard 0.213 to 0.0) while maintaining solid coherence (+0.138 C_v), as expected from clustering sentence embeddings with class-TF–IDF refinement [4]; the trade-off is a clear energy premium ($+4.03 \times 10^{-4}\,\mathrm{kWh}$), echoing Green AI cautions [6, 5]. Top2Vec underperforms on coherence (0.171 C_v) and draws more energy ($+6.6 \times 10^{-5}\,\mathrm{kWh}$) [14]. Finally, GSDMM shows undefined coherence and the highest energy, underscoring a mismatch with intra-line code-switching [15].

6 Conclusion

Our results point to a clear, workable path for topic modelling on short, code-switched Nigerian lyrics. Classical methods carry the day: NMF offers the strongest balance—high coherence, low overlap, tiny energy—while LSI provides a similarly light, interpretable baseline. Embedding pipelines shift the frontier differently: BERTopic delivers the crispest separation but at a notable energy premium; LDA remains frugal yet blurrier; HDP raises coherence while inflating redundancy; Top2Vec underperforms; and GSDMM is a poor fit for intra-line code-switching. In practice, start with NMF (or LSI when speed and simplicity are paramount), monitor diversity and overlap, and reserve BERTopic for cases that truly need strict topical disjointness and can afford the cost. This balance matters for classrooms, archives, and community projects running on limited power. Limitations include reliance on intrinsic metrics and energy, not emissions; future work will pair classical passes with targeted lightweight embeddings and incorporate expert, culturally grounded evaluations.

References

[1] Sakinat O Folorunso, Sulaimon A Afolabi, and Adeoye B Owodeyi. Dissecting the genre of nigerian music with machine learning models. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6266–6279, 2022.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[4] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based tf–idf procedure. *arXiv preprint*, 2022.

[5] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

[6] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13693–13696, 2020.

[7] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'Souza, Julia Kreutzer, Constantine Lignos, Salomey Osei, et al. MasakhaNER: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[8] F. Jayeola. Indigenous music in nigeria: Its role towards national development. *FUNAI Journal of Humanities and Social Sciences*, 1(2):102–109, 2015.

[9] Bode Omojola, editor. *Music and Social Dynamics in Nigeria*. Peter Lang Publishing, New York, NY, 2016.

[10] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[11] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[12] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[13] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM, 2015.

[14] Dimo Angelov. Top2Vec: Distributed representations of topics. *arXiv preprint*, 2020.

[15] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 233–242. ACM, 2014.

[16] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1456. ACM, 2013.

A Appendix A: Extended Experimental Details

A.1 Models and Training

We benchmark seven representative families of topic modeling approaches, chosen to capture the diversity of paradigms in the literature: (i) linear algebraic methods, (ii) probabilistic generative models, and (iii) recent embedding-based techniques.

This spread ensures that both classical baselines and modern neural-inspired models are evaluated under a common protocol. All experiments are conducted with a fixed random seed ($\texttt{seed} = 42$), and each configuration is repeated $R = 5$ times to account for variability. Unless otherwise noted, models are trained on CPU; GPU acceleration is employed only for methods that require deep embeddings.

LSI (Truncated SVD). Latent Semantic Indexing is implemented on TF–IDF representations with $k \in \{10, 20, 30, 40, 50\}$. Topics are derived from the highest loading terms in V_k:

$$X \approx U_k \Sigma_k V_k^\top.$$

NMF. Nonnegative Matrix Factorization is applied to TF–IDF features. Coordinate descent is used with $\texttt{beta\_loss} \in \{\textit{frobenius}, \textit{KL}\}$, $\alpha_W, \alpha_H \in \{0, 10^{-3}, 10^{-2}\}$, $\texttt{l1\_ratio} \in \{0, 0.2, 0.5\}$, and $\texttt{max\_iter} = 1000$:

$$X \approx WH, \quad W, H \geq 0.$$

LDA. Latent Dirichlet Allocation is trained on TF features using collapsed Gibbs sampling. We vary $\alpha \in \{0.1, 0.5, 1.0\}$ and $\eta \in \{1/V, 0.1\}$. Burn-in is 1000 iterations, followed by 2000 samples (thinning 10).

HDP. Hierarchical Dirichlet Process employs truncated stick-breaking at level 300. $\gamma = 1.0$ and $\alpha = 1.0$ are fixed, with 2000 training iterations. Low-mass topics are pruned post-training.

GSDMM (DMM). The Gibbs Sampling Dirichlet Multinomial Mixture follows a one-topic-per-document assumption, with $\alpha \in \{0.1, 0.5\}$ and $\beta \in \{0.1, 0.5\}$, running for 2000 sweeps.

BERTopic. Sentence embeddings are reduced with UMAP ($n\_neighbors \in \{5, 10, 15\}$, $\texttt{min\_dist} \in \{0.0, 0.1, 0.3\}$), then clustered with HDBSCAN ($\texttt{min\_cluster\_size} \in \{5, 10, 25\}$). Topics are represented using class-based TF–IDF.

Top2Vec. This approach jointly embeds documents and words in a shared semantic space. Topics are identified as dense regions; top words are extracted via nearest neighbor search around cluster centroids.

A.2 Metrics

For all evaluations, we report the mean $\pm 95\%$ confidence interval computed over R independent runs, using bootstrap resampling with 10,000 iterations. The following metrics are used:

Table 2: Summary of models, representations, and training procedures.

Model	**Representation**	**Training Method**	**Key Parameters**
LSI	TF–IDF	Truncated SVD	$k \in \{10, 20, 30, 40, 50\}$
NMF	TF–IDF	Coordinate descent, β-loss	$\alpha_W, \alpha_H \in \{0, 10^{-3}, 10^{-2}\}$; `l1_ratio` $\in \{0, 0.2, 0.5\}$; `max_iter` $= 1000$
LDA	TF	Collapsed Gibbs / VB	$\alpha \in \{0.1, 0.5, 1.0\}$; $\eta \in \{1/V, 0.1\}$; burn-in=1000; samples=2000
HDP	TF	Truncated stick-breaking	$\gamma = 1.0$, $\alpha = 1.0$; 2000 iters; truncation=300
GSDMM (DMM)	TF	Gibbs sampling	$\alpha \in \{0.1, 0.5\}$; $\beta \in \{0.1, 0.5\}$; 2000 sweeps
BERTopic	Sentence embeddings	UMAP $\rightarrow$ HDBSCAN $\rightarrow$ c-TF–IDF	$n\_neighbors \in \{5, 10, 15\}$; `min_dist` $\in \{0.0, 0.1, 0.3\}$; `min_cluster_size` $\in \{5, 10, 25\}$
Top2Vec	Doc+word embeddings	Dense clustering in embedding space	NN search over centroids

Table 3: Evaluation metrics with interpretation. Higher/lower arrows indicate preferred direction.

Metric	**Description**	**Interpretation**
C_{UMass}/C_v	Topic coherence (log co-occurrence and NPMI+cosine)	Higher $\uparrow$
Topic Diversity (TD)	Ratio of unique Top-N tokens to $k \cdot N$ total	Higher $\uparrow$
Inter-topic Overlap	Mean Jaccard similarity between topic word sets	Lower $\downarrow$
Indigenous Lexicon Coverage	Share of Top-N tokens in curated Yoruba/Pidgin lexicons	Higher $\uparrow$

B Appendix B: Additional Plots

This appendix provides extended visual analyses complementing the core results presented in the paper. These plots highlight the trade-offs between energy consumption, coherence, and diversity across classical and modern topic modelling approaches when applied to multilingual Nigerian music lyrics.

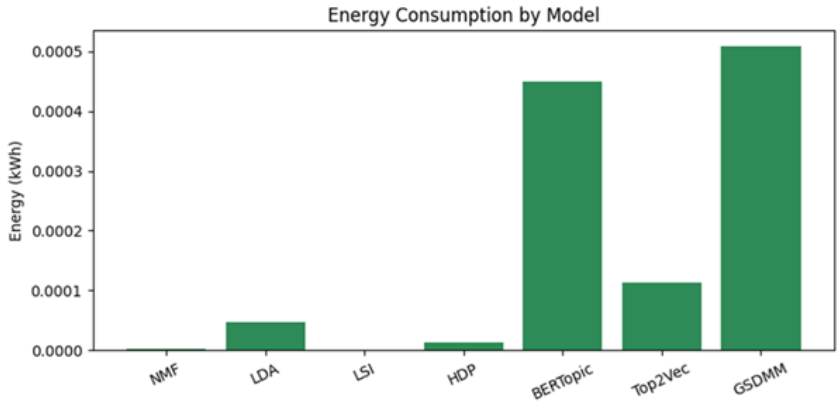

Figure 3: **Energy Consumption by Topic Modelling Approach.**

Figure 3 displays the energy consumption (kWh) of each topic modelling method when applied to the Nigerian music lyrics corpus. The bar chart provides a direct comparison of the computational sustainability of classical and advanced models. NMF, LDA, LSI, and HDP consume negligible energy, while BERTopic and GSDMM show the highest requirements, followed by Top2Vec. Transformer-based and clustering-intensive models, while innovative, require more energy—posing sustainability concerns for low-resource contexts. Classical models such as NMF and LSI are notably energy efficient, making them well-suited for carbon-conscious or resource-constrained deployments.

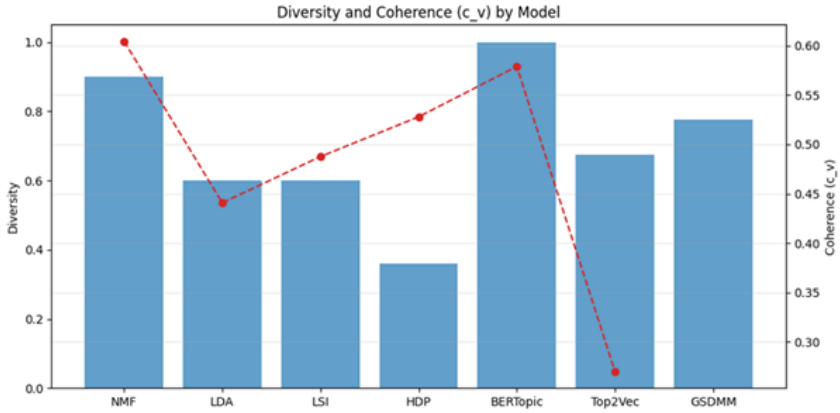

Figure 4: **Diversity and Coherence Across Topic Modelling Approaches.**

Figure 4 presents a dual-axis comparison of topic diversity (bars, left axis) and topic coherence C_v (red dashed line, right axis). BERTopic achieves the highest diversity and strong coherence, demonstrating its ability to discover distinct and meaningful topics. NMF offers the best overall coherence with high diversity. GSDMM and LSI perform moderately well, while HDP shows low diversity despite moderate coherence. LDA and Top2Vec underperform on both metrics. The figure highlights the importance of jointly optimizing diversity and coherence for cultural and multilingual corpora.

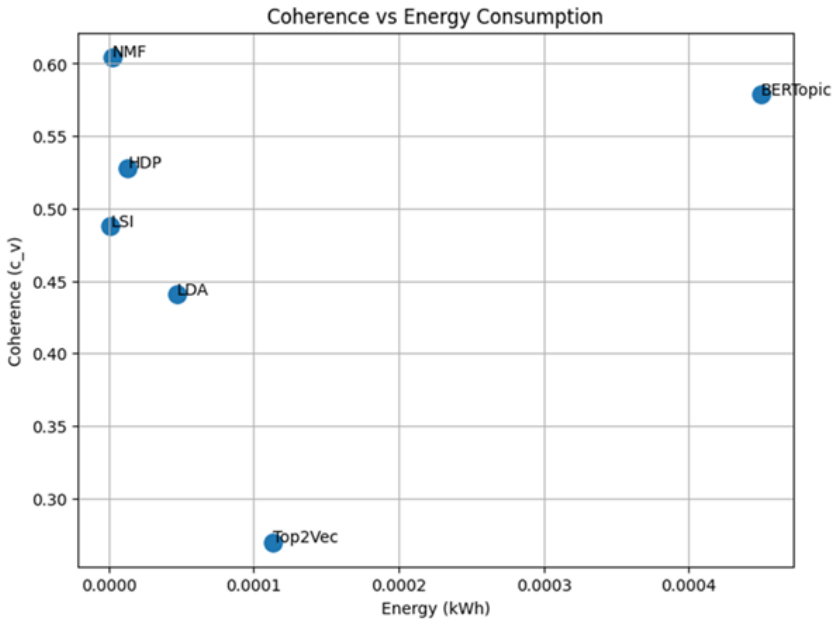

Figure 5: **Coherence versus Energy Consumption for Topic Modelling Approaches.**

Figure 5 illustrates the trade-off between semantic quality (coherence) and energy consumption. NMF stands out with the highest coherence and minimal energy usage. BERTopic also provides strong coherence but at a high energy cost, while LDA, LSI, and HDP strike an intermediate balance. Top2Vec records low coherence and higher energy use than most classical baselines. This figure underscores the importance of weighing both semantic performance and environmental impact in model selection for cultural text analysis.

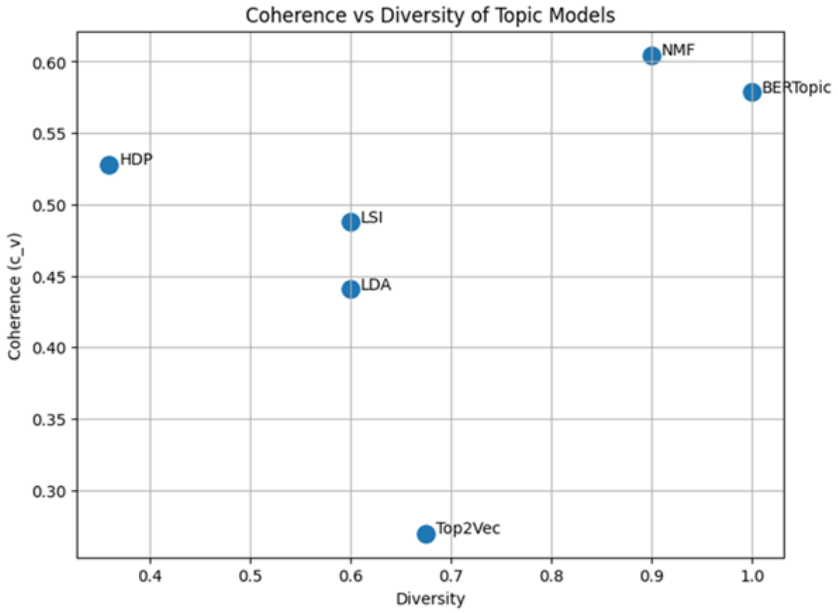

Figure 6: **Coherence versus Diversity of Topic Models.**

Figure 6 shows the relationship between topic coherence (C_v) and diversity across models. NMF leads in both coherence and diversity, closely followed by BERTopic. LDA and LSI provide moderate performance, while HDP shows limited diversity. Top2Vec performs weakest on coherence and only moderately on diversity. This trade-off visualization highlights how NMF and BERTopic provide optimal balance, making them promising for multilingual music analytics, while classical models remain competitive where computational efficiency is prioritized.