Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(RESEARCH)

Improving Public Service Grievance Analysis a Comparative Study of Topic Modelling Techniques with a Multi-Metric Data Cleaning Framework

Rahul Deka
National Informatics Centre, Assam, India

Reetesh Kumar Srivastava
National Informatics Centre, Assam, India

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2025, 2(3), 41–53

Article DOI: https://doi.org/10.70715/jitcai.2025.v2.i3.031

Abstract

The SewaSetu portal, a single window system for government services in Assam, India, processes thousands of applications and hundreds of grievances daily. And many such government grievance portals routinely receive a substantial volume of public complaints, each containing valuable information but often embedded in unstructured text. This study evaluates four topic modeling techniques—LSA, NMF, LDA, and BERTopic—on a large, anonymized grievance dataset processed through a multi-metric gibberish filtering pipeline. Model performance was assessed using coherence, topic diversity, and execution time across topic counts from 5 to 50. NMF consistently achieved the strongest coherence (0.7898 at K=35) while maintaining competitive diversity and efficiency. The results demonstrate that NMF, combined with robust preprocessing, provides a reliable framework for extracting interpretable themes from large-scale public grievance data and supports more informed service-delivery decision-making.

Keywords: Artificial Intelligence, BERTopic, E-Governance, Machine Learning, Natural Language Processing, Topic Modeling.

1. INTRODUCTION

Driven by the Digital India Programme, the government has prioritized the electronic delivery of services, resulting in significant initiatives to reform and simplify manual processes by establishing e-governance and offering various citizen services [1]. Sewasetu - a single window system for availing Government Services in Assam, offers 900+ services and processes more than 10,000 applications daily. While this has significantly streamlined service delivery, the portal's grievance redressal system faces a growing challenge. Currently, more than 100 grievances are received daily through various channels like the call center, email, and web forms. A public grievance officer manually consolidates and disburses the grievances to the concerned authority, which is a very time-consuming process. Looking at the growing nature of public complaints, as more and more e-governance services are getting on board into the portal, it would be humanly impossible beyond a point to manage the grievances. Therefore, the use of machine learning techniques in understanding and classifying grievances is proposed to automate the grievance redressal workflow. To automate and improve this process, this research explores the utilization of Topic Modeling for issue classification. Addressing the challenge of manually annotating massive datasets, this work focuses on first finding hidden topics within the authentic grievance datasets using a few existing models. Topic modeling offers a systematic way to interpret such unstructured text by grouping related ideas and uncovering overarching themes across large collections of complaints. Yet, topic-modeling methods vary considerably in how they operate; some perform well with traditional bag-of-words representations, while others rely on modern embedding-based techniques. Government complaint data further complicates this task, as entries are typically brief, unpredictable, and laden with domain-specific terminology.

This study adopts a practical, real-time data-driven approach by benchmarking several widely used topic-modeling methods, LDA, NMF, LSA, and BERTopic on a unified grievance dataset. We evaluate each method's ability to generate coherent and meaningful topics and identify which model is most reliable for real-world use by public agencies. The study on real-time data and Topic Modeling techniques promises to offer robust categorization of public grievances, which in turn can support decision-making and service improvement in both government and industrial applications.

2. RELATED WORK

Since citizen concerns related to service delivery are dynamic and continuously changing in nature, continuous exploration data analysis to understand and unearth the underlying issues through Topic Modeling becomes essential. The rapid growth of e-governance has resulted in a surge of citizen complaints through portals, call centers, and social media. Traditional manual grievance redressal systems are slow and inadequate, creating the need for automated solutions. Topic Modeling, an unsupervised NLP method, has shown strong potential for extracting hidden themes, classifying complaints, and supporting data-driven governance [2] [3]. Topic modeling has long been used in text mining, especially in public-sector analytics. Studies analyzing citizen complaints, urban service requests, and e-governance platforms often rely on LDA due to its simplicity. More recent work leans toward NMF and embedding-based approaches (like BERTopic), which tend to produce cleaner, more human-interpretable topics. Early studies relied on rule-based or keyword methods, but these failed to capture the diversity of language in complaints. Recent research demonstrated that models like LDA and NMF can automatically cluster complaints into coherent themes, covering issues such as loans, payments, and communication.

Recent research has increasingly focused on leveraging artificial intelligence and topic modeling techniques to enhance grievance analysis and e-governance systems. Sangeetha and Rao [4] applied Latent Dirichlet Allocation (LDA) to large-scale grievance datasets, demonstrating its effectiveness in identifying thematic structures. Building on this, Shah and Joshi [5], employed LDA on social media data for real-time complaint recognition, showcasing its adaptability to unstructured, dynamic environments. Das and Misra [6] integrated sentiment analysis into grievance redressal frameworks, improving decision support and prioritization of citizen concerns. Advancements in transformer-based architectures further enhanced classification accuracy and contextual understanding, as shown by Agarwal et al [7]. More recently, Gupta et al. [8] developed an AI-based solution enabling seamless grievance lodging and tracking across multiple government departments, reflecting a growing trend toward intelligent, citizen-centric e-governance platforms. These studies highlight both the scalability and interpretability of Topic Modeling in governance.

3. METHODOLOGY

Grievance portals are now the main way people submit complaints online, from slow services to problems with public infrastructure. Even though these platforms collect a huge amount of text, it’s still tough to understand the real issues because complaints can be messy, repetitive, and heavily context dependent. The dataset includes user-written complaints from an online grievance system, each describing a problem in free-form text. To make this information useful, we need computational methods that can automatically analyze, categorize, and prioritize the issues. The experimental pipeline is situated within standard grievance-management frameworks, wherein complaints move through the stages of capture, classification, routing, and redressal. According to the **NISG Service Maturity Model** and related digital governance frameworks, automated classification is a critical enablement layer for improving service responsiveness. Topic modeling in this work directly supports the classification stage by transforming unstructured grievances into structured thematic categories that can be integrated into public-service decision-support workflows.

3.1. DATA DESCRIPTION:

The original dataset comprised 91,866 records, of collected through a web form where both citizens and call centre operators registered grievances in the Sewasetu portal, which is a single window system for availing various government services in Assam state from September 2021 to November 2024. The initial dataset, gathered between September 2021 and November 2024, consists of 91,866 grievance records. These grievances were registered on the Sewasetu portal—Assam state's single-window system for government services—via a web form used by both citizens and call centre operators. The data includes short text complaints submitted that reflect real-world service issues encountered across departments. Each record contains unstructured text describing a service request, delay, or issue, making the dataset well-suited for evaluating topic modeling techniques on noisy, domain-specific content.

We hold the Intellectual Property Rights to all artefacts developed and produced under this project as per the Terms of Use. We retain full authority to modify, extend, or repurpose any technical components created as part of this work. We also reserve the right to claim awards or felicitation arising from these contributions and to publish technical papers or related outputs associated with this intellectual property.

DOI: https://doi.org/10.70715/jitcai.2025.v2.i3.031

3.2. DATA PRE-PROCESSING:

Standard NLP preprocessing steps were applied, including tokenization, lemmatization, removal of English stopwords, and vectorization using TF-IDF or CountVectorizer. To address the substantial noise present in real-world grievance text, a Multi-Metric Gibberish Filtering Pipeline was introduced. This pipeline incorporates structural integrity checks, lexical validity screening, entropy thresholds, perplexity-based randomness detection, and minimum-length constraints. Each metric targets a distinct noise pattern and operates independently to improve the reliability of the textual corpus before topic modeling. **U**nlike prior studies that rely primarily on lexical cleaning or stopword-based filtering, this work contributes a multi-metric gibberish detection framework specifically designed for government grievance data, representing a novel dimension of preprocessing in NLP research.

*3.2.1. **NER-Based Anonymization** To ensure compliance with the Digital Personal Data Protection (DPDP) Act, 2023, and to mitigate privacy risks when handling citizen-submitted grievances, an automated Named Entity Recognition (NER)–based anonymization pipeline was applied prior to any preprocessing or modeling [10][11]. This step specifically targets personal identifiers such as names and locations, which are among the most common forms of inadvertently shared sensitive information in public grievance data. The anonymization pipeline uses a RoBERTa-large model fine-tuned for NER to identify entities labeled PER (person) and LOC (location). For each text entry, the model detects named entities, and these are systematically replaced with standardized placeholders (e.g., <PER>, <LOC>) [12].*

*3.2.2. **Multi-Metric Gibberish Filtering** To ensure the reliability and interpretability of the extracted topics, a rigorous data-cleaning phase was implemented with emphasis on identifying and removing low-quality or "gibberish" entries commonly present in unstructured grievance text. A Multi-Metric Gibberish Scoring Pipeline was developed to assign each record a composite Gibberish Score S_g, normalized between 0 (clean) and 1 (gibberish), computed as the weighted sum of five normalized metrics: structural integrity, lexical validity, Shannon entropy [13], perplexity, and minimum-length violation. These metrics capture distinct noise patterns, ranging from malformed character sequences to statistically random or semantically incoherent text.*

Equal weights (0.2 each) were used intentionally because no single metric dominates empirically; preprocessing logs showed that each metric flagged a largely disjoint subset of noisy entries, demonstrating complementary value. The threshold of 0.3 was selected after iterative inspection of borderline cases, ensuring that meaningful but short or informal grievances were retained while eliminating syntactically corrupt or nonsensical submissions. Although a full ablation study is beyond the current scope, preliminary tests indicated that removing any one metric reintroduced specific classes of noise, reaffirming the need for a multi-metric composite score.

A summary of all five metrics, their purpose, and Metric calculation mechanism is provided in *Table 1* for clarity and reproducibility.

$$S_g = \sum_{i=1}^{5} W_i \times N(M_i)$$

S_g= Gibberish score

W_i= Weight (set as 0.2)

Mi=value of i^{th} metric

$N(M_i)$= Normalization Function (N) standardizes i^{th} Metric M_i

The Calculation of S_g can be broken down into the following four steps.

I. Compute Each Raw Metric M_i (Refer to Column Metric calculation in Table 1)
II. Apply Normalization to Each Metric
III. Apply the Fixed Weight (0.2 Each)
IV. Applying the Threshold

Table 1: Summary of metrics and their calculation method used in the Multi-Metric Gibberish Filtering Pipeline

Metric	Description/ Purpose	Metric Calculation
Structural Integrity	Measures ratio of non-alphabetic characters to text length	$M_1 = \frac{Non-alphabetic\ characters}{Total\ characters}$
Lexical Validity	Checks presence of dictionary-valid tokens	$M_2 = 1 - \frac{valid\ dictionary\ tokens}{Total\ tokens}$
Shanon Entropy	Detects randomness or excessive repetition	M_3 =H =-$\sum p(x) log_2 p(x)$
Perplexity	Measures linguistic probability of the sequence	M_4= 2^H
Minimum-Length Constraint	Flags texts shorter than meaningful thresholds	$M_5 = \{\frac{1, if\ text\ length < threshold}{0, Otherwise}$

Apply Normalization to Each Metric- Because metrics operate on different scales, each raw metric M_i is passed through a normalization function which converts all metrics to the same 0–1 scale, ensuring fairness.

$$N(M_i) = \frac{M_i - min(M_i)}{max(M_i) - min(M_i)}$$

Apply the Fixed Weight (0.2 Each)-Since all weights Wi=0.2.

$$S_g = 0.2N(M_1) + 0.2N(M_2) + 0.2N(M_3) + 0.2N(M_4) + 0.2N(M_5)$$

The above equation is equivalent to $S_g = \sum_{i=1}^{5} W_i \times N(M_i)$

Apply the Threshold- S_g>0.3

This threshold was selected empirically, ensuring: Minor noise does not remove valid grievances, and truly malformed or incoherent texts are filtered out

*3.2.3. **Deduplication:** The dataset was examined for identical grievances, and duplicate rows based on the Cleaned Text content were removed. This step ensured that the statistical weight of each unique grievance was represented only once in the subsequent topic modeling phase.*

*3.2.4. **Text Normalization** Following the corpus cleaning phase, the remaining text documents were prepared for topic modeling through standard text normalization steps:*

Basic Cleaning: All text is lowercased, and punctuation, digits, and special characters are removed to standardize the input.

Tokenization: The text is split into individual tokens (words) [14].

Stop-word Removal: Common words that carry little semantic value (e.g., “the”, “is”, “and”) are removed using standard stopword lists [15].

Lemmatization: Tokens are reduced to their base or dictionary form (e.g., “running” → “run”) to normalize variations of the same word [16].

Ultimately, preprocessing plays a central role in enabling scalable, accurate, and efficient grievance redressal in e-governance. Ensuring consistency and quality at the earliest stage sets the stage for intelligent systems that can support both operational efficiency and policymaking.

3.3. Document Representation

Different topic modeling algorithms depend on different mathematical formulations; the study constructed separate document–term representations for each model category. For LDA, a bag-of-words representation was created using a count-based document–term matrix, as this form aligns with its probabilistic generative assumptions [17]. For NMF and LSA, a TF–IDF matrix was generated to capture the relative importance of words across the corpus and to mitigate the influence of frequently occurring, low-information terms [18].

For BERTopic, document-level semantic embeddings were generated using the "all-MiniLM-L6-v2" Sentence Transformer, producing 384-dimensional dense vectors. These embeddings serve as the input to UMAP for dimensionality reduction and HDBSCAN for density-based clustering, enabling BERTopic to identify semantic patterns that extend beyond surface-level token overlap [19] (*parameters associated with these representations are provided in Table 2*). Together, these representation choices ensure that each algorithm receives inputs tailored to its underlying assumptions while maintaining methodological consistency across the experimental pipeline.

3.4. Topic Modeling Algorithms

Non-Negative Matrix Factorization (NMF) relied on factorizing the TF–IDF matrix into non-negative document–topic and topic–word matrices. The model approximated the input matrix using additive components, making it well-suited to short, information-dense text such as complaints [20]. Latent Dirichlet Allocation (LDA) was implemented using the Gensim framework, employing a variational Bayes optimization scheme [21]. The model treated each document as a mixture of latent topics, with each topic characterized by a probability distribution over words [22]. Latent Semantic Analysis (LSA) applied truncated singular value decomposition to the TF–IDF matrix. Although traditionally known for capturing latent semantic structure, it served primarily as a comparative baseline due to its tendency to mix heterogeneous concepts [23]. BERTopic employed a modern, embedding-driven pipeline. Documents were encoded using a transformer model, projected into a lower-dimensional space with UMAP, and clustered using HDBSCAN [24]. Topics were then extracted using a class-based TF–IDF procedure that highlights words most representative of each cluster [25] [26].

All models were trained on the same data, and the number of topics was fixed to 20 for the primary comparison, except for BERTopic, which determines topic quantity based on cluster density. ***Fig 1*** shows a conceptual pipeline showing the end-to-end workflow from citizen grievance collection through text preprocessing and topic modeling to actionable insights. clearly illustrates the mechanism of the experiment from Data Pre-processing to benchmarking and evaluation of topic coherence.

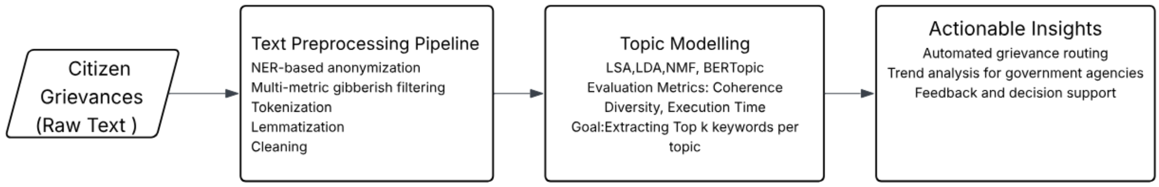
Citizen Grievances (Raw Text)
Text Preprocessing Pipeline
NER-based anonymization
Multi-metric gibberish filtering
Tokenization
Lemmatization
Cleaning
Topic Modelling
LSA,LDA,NMF, BERTopic
Evaluation Metrics: Coherence Diversity, Execution Time
Goal:Extracting Top k keywords per topic
Actionable Insights
Automated grievance routing
Trend analysis for government agencies
Feedback and decision support

Fig 1: Conceptual pipeline showing the end-to-end workflow

3.5. Experimental Configuration (Parameters)

A consistent set of preprocessing, vectorization, and model-specific parameters was applied across all four topic modeling algorithms to ensure a fair comparison. NMF was trained with n_components = K; random_state = 42; max_iter = 200 . LDA used n_components = K, random_state = 42, and n_jobs = -1 to leverage parallel execution. BERTopic was implemented with the "all-MiniLM-L6-v2" Sentence Transformer model, combined with a UMAP reducer configured as n_neighbors = 15, n_components = 5, min_dist = 0.0, metric = "cosine", and an HDBSCAN clustering stage using the default density-based settings.

All vectorization and model-specific parameter settings used in this study are summarized in *Table 2* to ensure transparency and reproducibility.

All experiments were conducted on Python 3.10 using scikit-learn 1.3, Gensim 4.3, and BERTopic 0.15.2. Sentence embeddings were computed with SentenceTransformers 2.2.2. Execution was performed on a system with an Intel i7 CPU, 16 GB RAM, and an NVIDIA T4 GPU for models requiring acceleration.

DOI: https://doi.org/10.70715/jitcai.2025.v2.i3.031

Table 2: Parameter setting for various components used in this experiment

Component	Parameter Setting
Topic Range	Number of topics K={5,10,15,20,25,30,35,40,45,50}
Top Words per Topic	10 words (for coherence, diversity);
TF-IDF Vectorization	max_df=0.95: Terms appearing in more than 95% of documents were ignored.
LSA	n_components: Set to the varying random_state=42:
NMF	n_components = K; random_state = 42; max_iter = 200
LDA	n_components: Set to the varying num_topics. random_state=42: For reproducibility. n_jobs=-1: Utilizes all available CPU cores for faster processing.

BERTopic	nr_topics = {5–50}, embedding_model = "all-MiniLM-L6-v2", calculate_probabilities = False,
	verbose = False,

BERTopic UMAP	n_neighbors = 15; n_components = 5; min_dist = 0.0; metric = "cosine"
BERTopic HDBSCAN	Default BERTopic/HDBSCAN settings
Gensim Dictionary & Corpus (for Coherence Calculation)	no_below=2 (filter words that appear in fewer than 2 documents) and no_above=0.95 (filter words that appear in more than 95% of documents

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

To determine the optimal topic modeling configuration for the grievance dataset, a systematic benchmarking study was performed across four established algorithms: NMF, BERTopic, LDA, and LSA. The primary experimental variable was the number of topics (K). By assessing model performance across multiple topic ranges (K = 5–25, 5–40, and 5–50) using three complementary metrics—coherence, topic diversity, and execution time—the study provides a comprehensive evaluation of both semantic quality and computational efficiency across all four algorithms.

Model performance was assessed using three complementary evaluation measures: topic coherence (C_v), topic diversity, and execution time, enabling a balanced assessment of semantic quality, inter-topic distinctiveness, and computational efficiency. The overall experimental workflow used in this study is depicted in *Fig. 2*.

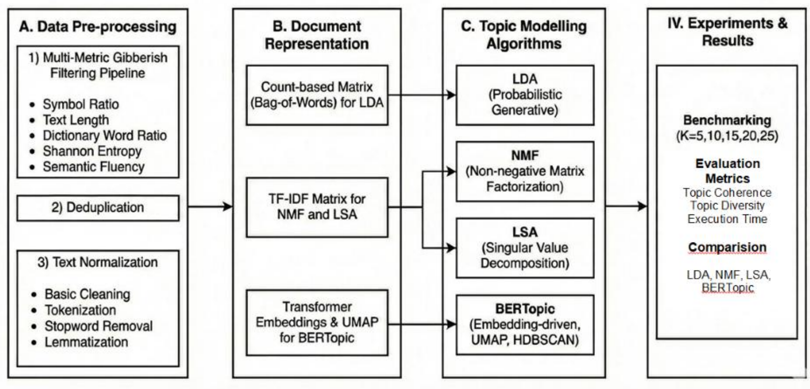
A. Data Pre-processing
1) Multi-Metric Gibberish Filtering Pipeline
• Symbol Ratio
• Text Length
• Dictionary Word Ratio
• Shannon Entropy
• Semantic Fluency
2) Deduplication
3) Text Normalization
• Basic Cleaning
• Tokenization
• Stopword Removal
• Lemmatization
B. Document Representation
Count-based Matrix (Bag-of-Words) for LDA
TF-IDF Matrix for NMF and LSA
Transformer Embeddings & UMAP for BERTopic
C. Topic Modelling Algorithms
LDA (Probabilistic Generative)
NMF (Non-negative Matrix Factorization)
LSA (Singular Value Decomposition)
BERTopic (Embedding-driven, UMAP, HDBSCAN)
IV. Experiments & Results
Benchmarking (K=5,10,15,20,25)
Evaluation Metrics
Topic Coherence
Topic Diversity
Execution Time
Comparision
LDA, NMF, LSA, BERTopic

Fig 2: Flowchart of the Experiment (Data Pre-processing to Result)

*4.1.1. **Coherence Score:** The C_v coherence score, computed using the standard Gensim implementation [27], was used to assess intra-topic semantic quality. Higher C_v values reflect stronger semantic relatedness among the top words within each topic, thereby indicating more interpretable and meaningful topic structures. Standard implementations of LSA, NMF, and LDA were used for coherence evaluation, while BERTopic leveraged precomputed sentence embeddings from the SentenceTransformer model "all-MiniLM-L6-v2" to support stable and efficient topic formation across the evaluated K values. [28].*

*4.1.2. **Topic diversity** was used as a complementary measure of topic quality, capturing the degree of lexical uniqueness across topics. While coherence assesses intra-topic semantic consistency, high coherence alone may mask redundancy if multiple topics share similar vocabulary. Topic diversity, therefore, evaluates inter-topic distinctiveness by computing the ratio of unique words to the total number of words in the top ten terms extracted from all topics (as shown in the expression below). A score of 1.0 indicates that all top words are unique, whereas lower values reflect increasing overlap and reduced thematic separation. [29].*

$$Topic\ Diversity = \frac{Unique(W)}{T \times k}$$

Unique(W)= number of distinct words across all topics

T x k= Total number of words used in diversity calculation

*4.1.3. **Execution time** was recorded to evaluate the computational efficiency of each algorithm. For every K, the total time taken by the model to fit and generate topics was measured. This metric provides insight into practical usability, especially for large-scale, real-time, or resource-constrained e-governance applications where model speed can become a critical operational factor.*

4.2. Results:

The performance of the four topic modeling algorithms was evaluated using three key parameters: topic coherence, topic diversity, and execution time to provide a comprehensive assessment of their semantic quality, distinctiveness, and computational efficiency.

*4.2.1. **Coherence Score-** NMF showed the most stable and increasing coherence trend as the number of topics increased, indicating strong semantic grouping of grievance themes. LDA performed moderately well at lower topic counts but produced increasingly generic topics at higher K values. BERTopic exhibited fluctuations due to its density-based clustering behavior, while LSA consistently underperformed because its linear decomposition struggles with short, noisy text.The combined results across all three experiments consistently identify NMF as the most coherent and scalable approach, capable of producing interpretable and operationally meaningful grievance topics even at higher topic counts. **Fig. 3** visualizes the trend of C_v (coherence score) across K=25, K=40, and K=50, respectively.*

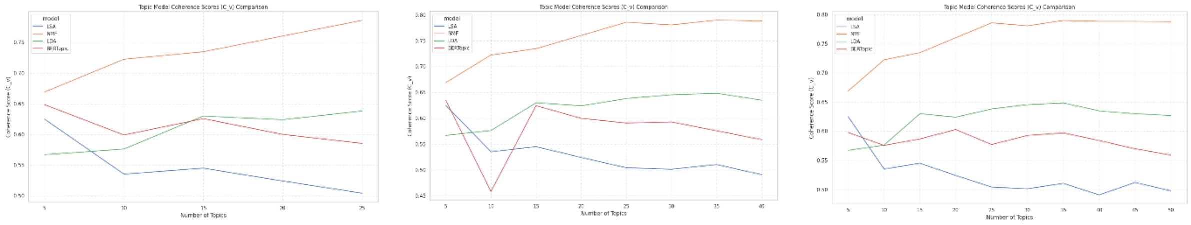

Fig 3: Coherence Score (Cv) trend for four models at K=5 to K=50

*4.2.2. **Topic diversity** computed using the top 10 words per topic shows a clear decreasing trend as the number of topics increases across all four models. At K = 25, NMF maintains the strongest lexical distinctiveness (≈0.85), indicating minimal word overlap and well-separated themes. BERTopic performs moderately well (≈0.81), followed by LDA with diversity values around 0.63–0.64. LSA exhibits the weakest performance (≈0.39), reflecting substantial redundancy in its topic-word distributions.As the topic count increases to K = 40 and K = 50, the performance gap between models becomes more pronounced. NMF continues to demonstrate high diversity (≈0.83 at K = 40 and ≈0.80 at K = 50), showing strong scalability and the ability to preserve distinct themes even under fine-grained topic settings. In contrast, BERTopic's diversity declines to approximately 0.68–0.70 due to denser clustering in the embedding space, while LDA remains moderate at around 0.63–0.65. LSA consistently yields the lowest diversity scores (≈0.35), indicating limited capability to preserve separable topic structures. Overall, NMF emerges as the only model capable of sustaining meaningful topic separation as K increases. **Fig. 4** illustrates Topic diversity trends for four models from K = 5 to 50, with NMF maintaining the highest uniqueness of top words.*

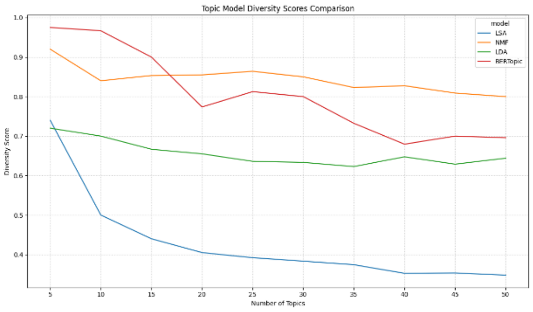

Fig. 4: Topic diversity trends for four models from K = 5 to 50

*4.2.3. **Execution time** represents the total duration required for each algorithm to fit the topic model at different values of K. Execution times increased with topic count for all models, with LDA being the slowest because of iterative inference, NMF remaining moderate, and LSA performing fastest due to its linear decomposition structure. BERTopic showed higher time consumption because of UMAP and HDBSCAN clustering overhead. **Fig 5** illustrates the execution time comparison of four topic modeling algorithms across varying topic counts (K = 5–50).*

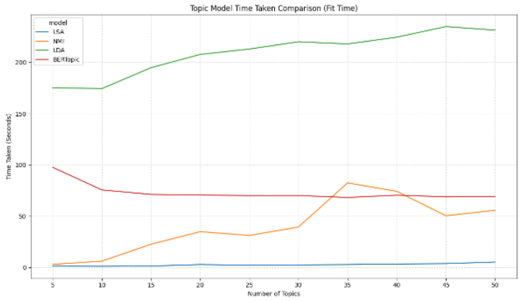

Fig 5: Execution time comparison of four topic modelling algorithms across varying topic counts (K = 5–50).

Table 3 shows a comparative performance summary of four topic modeling algorithms (NMF, BERTopic, LDA, and LSA) based on coherence trends, diversity, execution time, and observed behavioral patterns.

Table 3: Comparative performance Summary of NMF, BERTopic, LDA, and LSA

Model	Coherence Trend Description	Highest Cv Coherence Score	Diversity at K=50	Execution time (in seconds)	Behavioural Observation
NMF	Strong, consistently increasing trend across all ranges (5→25→40→50). Coherence rises smoothly and stabilizes near the upper bound.	≈ 0.79 ($K \approx 30 - 40$)	0.8000	2.17	Stable and consistent; high coherence and diversity with low execution time.
BERTopic	Fluctuating pattern. Peaks around K = 10–15, then dips, then mild recovery, but steadily declines after K 30.	$\approx$ 0.63 (K = 15)	0.6959	151.47	Fluctuating coherence; moderate diversity; high computation cost.
LDA	Gradually increasing until mid-range K (15–35), then stabilizes with slight fluctuations.	$\approx$ 0.65 (K = 35)	0.6440	142.78	Gradual improvement then stabilizes; moderate coherence and diversity; slow.
LSA	Sharp decline from K = 5 to 10, then slight oscillation with overall decreasing trend as K increases.	$\approx$ 0.625 (K = 5)	0.3480	0.31	Sharp coherence drops; low diversity; fast but poor topic quality.

A deeper qualitative analysis of representative topics from each model is detailed in ***Table 4***, highlighting the varying levels of specificity and coherence achieved across the algorithms.

Table 4: Qualitative Analysis of Topic Keywords

Model	Topic ID (K-Setting)	Top Keywords	Analysis
NMF	Topic 0	*caste caste certificate* *apply caste apply rtps caste issuance caste month issuance receive caste*	Excellent Coherence & Specificity. Highly focused on the *Caste Certificate Issuance* lifecycle. The clean separation supports NMF's high C_v score at $K = 35$.
BERTopic	Topic 0	*certificates apply application rtps* *caste date delivers sir land process*	Moderate Coherence. Covers general apply- cation steps but blends distinct services (*caste, land*). The inclusion of generic words (*sir*) indicates a broader, less-specific thematic cluster.
LDA	Topic 0	*copy sir apply madam certify sir* *madam dear jamabandi certify copy dear sir*	Poor Coherence (Junk Topic). Dominated by generic salutations and bureaucratic template language. Fails to capture a substantive service issue, demonstrating a common failure mode of LDA at high *K*.
LSA	Topic 0	*certificate caste certificate* *apply application rtps date deliver apply caste delivery*	Coherent but General. Functions as a primary theme, encompassing the dominant *Certificate Delivery and Application* services. Lacks the granular specificity achieved by the optimal NMF configuration.

5. Algorithmic Discussion and Critical Comparison

5.1. NMF's Mathematical Superiority for Short Text Documents-

The strong, consistent performance of NMF, which outperformed alternatives across all tested ranges and stabilized with high coherence scores near 0.79, is directly related to its algebraic foundation. NMF's core mechanism involves factorizing the TF-IDF matrix into non-negative document-topic and topic-word matrices. The technical advantage of NMF over probabilistic models like LDA, especially when handling short, sparse text like public complaints, is its reliance on non-negative, additive components. Grievance data, characterized by short length and few word co-occurrences, leads to highly sparse document-term matrices. While LDA struggles to infer reliable probabilistic structures in minimal co-occurrence contexts, NMF's algebraic approach effectively models word strength (weight) directly from the TF-IDF matrix. This makes NMF significantly more robust and reliable for discovering specific, information-dense concepts within sparse short texts.

5.2. Analysis of Algorithm Deterioration (LSA and LDA)

The evaluation reveals a stark divergence in performance between Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) when applied to short, noisy grievance texts. LSA exhibits a sharp, consistent decline in topic coherence as the number of topics K exceeds 5. This deterioration stems from LSA's reliance on Singular Value Decomposition (SVD) of a sparse document-term matrix. With brief documents, co-occurrence data is limited, causing SVD to merge heterogeneous concepts into latent dimensions. Increasing K merely fragments these mixed dimensions, amplifying instability and incoherence; a clear indication that LSA is ill-suited for fine-grained topic extraction in such contexts.

LDA exhibits instability at higher topic counts because its probabilistic generative model requires sufficient word co-occurrence evidence to estimate topic–word distributions. Grievance texts are short and sparse, resulting in weak statistical support as K increases. This forces LDA to assign probability mass to high frequency but semantically uninformative terms (e.g., "sir", "apply", "madam"), producing generic or "junk" topics at higher K.

5.3. Interpretation of BERTopic Fluctuations

BERTopic demonstrated moderate performance, peaking in the mid-range of K (around K=10-15), but showing declining and fluctuating performance as topic counts exceeded K=30. BERTopic's fluctuations arise from the interaction between UMAP and HDBSCAN. UMAP performs stochastic manifold reduction, and HDBSCAN forms clusters based on variable density regions. At larger K values, the embedding space becomes increasingly fragmented; clusters become smaller and may sit in low-density regions, causing instability in topic formation.

6. CONCLUSION

This study presented a comprehensive evaluation of four topic modeling algorithms—LSA, NMF, LDA, and BERTopic—on a large-scale government grievance dataset, incorporating a robust preprocessing pipeline that included *NER-based anonymization* to ensure compliance with the *Digital Personal Data Protection (DPDP) Act, 2023,* and multi-metric gibberish filtering. By assessing model performance across multiple topic ranges (K = 5–25, 5–40, and 5–50) and three complementary metrics: coherence, topic diversity, and execution time. The analysis offers a holistic understanding of each model's strengths and limitations. The findings consistently demonstrate that NMF achieves the highest semantic coherence and maintains strong topic diversity even at higher topic counts, while also exhibiting reasonable computational efficiency. In contrast, LDA and BERTopic show moderate but less stable performance across K values, and LSA proves unsuitable for fine-grained topic separation due to low diversity and high redundancy. The technical advantage of NMF over probabilistic models like LDA, especially when handling short, sparse text like public complaints, is its reliance on non-negative, additive components. Grievance data, characterized by short length and few word co-occurrences, leads to highly sparse document-term matrices. While LDA struggles to infer reliable probabilistic structures in minimal co-occurrence contexts, NMF's algebraic approach effectively models word strength (weight) directly from the TF-IDF matrix. This makes NMF significantly more robust and reliable for discovering specific, information-dense concepts within sparse short texts

Overall, the results establish NMF as the most reliable and scalable method for modeling grievance narratives in e-governance environments. Its ability to preserve interpretability at varying granularities makes it particularly well-suited for operational deployments such as automated grievance routing, issue trend detection, and policy monitoring. Future work may extend this framework by incorporating human evaluation, exploring multilingual grievance streams, and integrating real-time topic drift detection to support dynamic governance ecosystems.

7. FUTURE DIRECTION

With this strong foundation in both robust NMF and coherence benchmarking, the next step in this research will incorporate large language models to further advance topic interpretation and analysis, moving toward LLM-Augmented Topic Modeling. We will use these models to automate the summarization and categorization of topics, having LLMs generate concise, human-readable labels and high-level summaries from the most weighted keywords. Second, interpretability-driven optimization will be explored by utilizing LLMs to assess the semantic quality and distinctiveness of topics, providing a more human-centered metric for honing in on the optimal number of topics, K, beyond what is feasible with the C_v coherence metric. Finally, we will explore hybrid topic generation by integrating large language models into state-of-the-art models like BERTopic, refining term weights at the core structure to maximize coherence and separation in challenging, overlapping topics, and yield a more efficient and scalable large-scale public grievance data analytical framework.

8. References

[1] K. Chaudhury, A. Barua, R. Deka, T. Gogoi, A. C. Gupta, and S. Pyarelal, "Reforming and Strengthening Digital Service Delivery: Case of Government of Assam.," in *National Conference for e-Governance, Mumbai, India. 2020.*, Jun. 2020.

[2] Z. Tang, X. Pan, and Z. Gu, "Analyzing public demands on China's online government inquiry platform: A BERTopic-Based topic modeling study," *PLOS ONE*, vol. 19, no. 2, p. e0296855, Feb. 2024, doi: 10.1371/journal.pone.0296855.

[3] I. Spasic and G. Nenadic, "Clinical Text Data in Machine Learning: Systematic Review," *JMIR Medical Informatics*, vol. 8, no. 3, p. e17984, Mar. 2020, doi: 10.2196/17984.

[4] S. G and L. M. Rao, "Modelling of E-Governance Framework for Mining Knowledge from Massive Grievance Redressal Data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 1, p. 367, Feb. 2016, doi: 10.11591/ijece.v6i1.pp367-374.

[5] K. Shah, and H. Joshi, "Smart approach to recognize public grievance from microblogs," Towar. Excell. UGC HRDC GU, vol. 13, no. 02, pp. 57–69, 2021, doi: 10.37867/te130206.

[6] R. K. Das, M. Panda, and H. Misra, “Decision support grievance redressal system using sentence sentiment analysis,” in *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, New York, NY, USA: ACM, Sep. 2020, pp. 17–24. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1145/3428502.3428505

[7] S. Agarwal and A. Sureka, “Investigating the Role of Twitter in E-Governance by Extracting Information on Citizen Complaints and Grievances Reports,” in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2017, pp. 300–310. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1007/978-3-319-72413-3_21

[8] P. Gupta, O. P. Ijardar, A. Jadhav, and V. Saheb, “AI-Based Solution To Enable Ease of Grievance Lodging and Tracking for Citizens Across Multiple Departments,” in *Advances in Intelligent Systems Research*, Dordrecht: Atlantis Press International BV, 2025, pp. 1002–1022. Accessed: Dec. 04, 2025. [Online]. Available: https://doi.org/10.2991/978-94-6463-738-0_78

[9] SS. Vijayarani, M. J. Ilamathi, M. Nithya, et al., “Preprocessing techniques for text mining-an overview,” International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 716–, 2015.

[10] S. Sharma, R. K. Srivastava, and P. P. Singh, “Advancements in Named Entity Recognition using Deep Learning Techniques: A Comprehensive Study on Emerging Trends,” *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 5, pp. 256–262, Jun. 2023, doi: DOI: 10.48175/IJARSCT-11657.

[11] B. Mohit, “Named Entity Recognition,” in *Theory and Applications of Natural Language Processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 221–245. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1007/978-3-642-45358-8_7

[12] Y. Sun, D. Gao, X. Shen, M. Li, J. Nan, and W. Zhang, “Multi-Label Classification in Patient-Doctor Dialogues With the RoBERTa-WWM-ext + CNN (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach With Whole Word Masking Extended Combining a Convolutional Neural Network) Model: Named Entity Study (Preprint),” JMIR Publications Inc., Dec. 2021. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.2196/preprints.35606

[13] D. Ellerman, “The Relationship Between Logical Entropy and Shannon Entropy,” in *SpringerBriefs in Philosophy*, Cham: Springer International Publishing, 2021, pp. 15–22. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1007/978-3-030-86552-8_2

[14] D. Yogish, T. N. Manjunath, and R. S. Hegadi, “Review on Natural Language Processing Trends and Techniques Using NLTK,” in *Communications in Computer and Information Science*, Singapore: Springer Singapore, 2019, pp. 589–606. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1007/978-981-13-9187-3_53

[15] K. Ghosh and A. Bhattacharya, “Stopword Removal,” in *Proceedings of the 10th Annual ACM India Compute Conference*, New York, NY, USA: ACM, Nov. 2017, pp. 99–102. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1145/3140107.3140125

[16] C. Landau, “Understanding Stemming and Lemmatization,” in *Mastering Natural Language Processing Part 2*, Berkeley, CA: Apress, 2024. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1007/979-8-8688-0549-3_1

[17] W. A. Qader, M. M. Ameen, and B. I. Ahmed, “An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges,” in *2019 International Engineering Conference (IEC)*, IEEE, Jun. 2019, pp. 200–204. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1109/iec47844.2019.8950616

[18] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[19] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *Journal of Open Source Software*, vol. 3, no. 29, p. 861, Sep. 2018, doi: 10.21105/joss.00861.

[20] “Non-negative Matrix Factorization,” in *Definitions*, Qeios, 2020. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.32388/b2qmcz

[21] B. Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” in *Advances in Neural Information Processing Systems 14*, The MIT Press, 2002, pp. 601–608. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.7551/mitpress/1120.003.0082

[23] N. E. Evangelopoulos, “Latent semantic analysis,” *WIREs Cognitive Science*, vol. 4, no. 6, pp. 683–692, Aug. 2013, doi: 10.1002/wcs.1254.

[24] G. Stewart and M. Al-Khassaweneh, “An Implementation of the HDBSCAN* Clustering Algorithm,” *Applied Sciences*, vol. 12, no. 5, p. 2405, Feb. 2022, doi: 10.3390/app12052405.

[25] R. K. Srivastava, S. Sharma, and P. P. Singh, “Exploring Latent Themes-Analysis of Various Topic Modeling Algorithms,” *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 5, pp. 225–229, Jun. 2023, doi: DOI: 10.48175/IJARSCT-11635.

[26] J. L. Rachel J, B. A, and K. M, “Topic Modeling Based Clustering of Disaster Tweets Using BERTopic,” in *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*, IEEE, Apr. 2024, pp. 1–6. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1109/mitadtsocicon60330.2024.10575555

[27] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, Feb. 2015, pp. 399–408. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[28] C. Yin and Z. Zhang, “A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With ‘Same Semantics, Different Structure’ After Fine Tuning,” in *Advances in Computer Science Research*, Dordrecht: Atlantis Press International BV, 2024, pp. 677–684. Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.2991/978-94-6463-540-9_69

[29] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” *arXiv:2203.05794*, Mar. 2022, doi: https://doi.org/10.48550/arXiv.2203.05794 Focus to learn more.