Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 08 Issue 02
2025

Research Article
https://doi.org/10.56979/802/2025

Topic Modeling Empowered by a Deep Learning Framework Integrating BERTopic, XLM-R, and GPT

Nooria Aamir[1], Ali Raza[1], Muhammad Waseem Iqbal[1,] Khalid Hamid[1,*], Zaeem Nazir[1], Ayyan Asif[2], Samia Hussain[3], and Hafiz Abdul Basit Muhammad[1]

[1]Department of Computer Science and Information Technology, Superior University Lahore, Lahore, 54000, Pakistan
[2]Master of Science in Data Analytics (Stem) Department of Computer Science New Mexico State University, Las Cruces, NM
[3]Department of Computer Science, UET Lahore, 54000, Pakistan
[*]Corresponding Author: Muhammad Waseem Iqbal. Email: waseem.iqbal@superior.edu.pk

Received: January 29, 2025 **Accepted:** March 01, 2025

Abstract: Topic modeling facilitates the identification of hidden themes and patterns in large text collections. It enables a thorough investigation of the messages contained in texts. Topic modelling is a popular research subject, with several translations already being investigated, including English and Arabic. However, there is a need for more research into low-resource languages, including Urdu. In this study, we propose using the BERTopic, XLM-R, and GPT frameworks on Urdu text. The proposed approach, which includes fine-tuned BERT, XLM-R, and GPT models, aims to capture the contextual nuances and grammatical intricacies of Urdu text. In this investigation, we used existing Urdu textual data. We evaluated the performance of our proposed approaches to existing techniques such as LDA and NMF utilizing coherence and diversity measures. The results show that our proposed strategy outperforms existing methods, with an average coherence improvement of 0.05 and a diversity score of 0.87. These findings demonstrate the efficacy of the proposed approach in extracting significant topics from Urdu texts, hence assisting scholarly endeavors in comparative studies of Urdu translations. Integrating real-time Urdu topic modeling into social media and news monitoring systems can help in trend analysis, misinformation detection, and sentiment-aware content moderation. Another practical application is the incorporation of topic modeling in Urdu search engines and recommendation systems, improving information retrieval for Urdu-speaking users.

Keywords: Topic Modeling; Language modeling; XLM-R; GPT

1. Introduction

Topic modeling is a widely utilized technique in natural language processing (NLP) for discovering hidden thematic structures within large collections of text data. By automatically identifying latent topics within textual corpora, topic modeling facilitates the organization of documents, providing valuable insights into the content [1, 2]. Among the most prominent methods are Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), which have been successfully applied to high-resource languages such as English, Arabic, and Chinese [1, 3]. These methods have enabled significant advancements in various fields, including information retrieval, content recommendation, and summarization [4, 5].

However, while these techniques have been well-explored for high-resource languages, a notable gap persists in their application to low-resource languages. This gap is particularly evident in languages such as Urdu, which remains underrepresented in the field of computational linguistics. Urdu is a morphologically rich language with complex syntactic structures, and its linguistic diversity, including phenomena such as code-switching and Urdu-Hindi dialectal variation, further complicates its analysis using conventional NLP techniques [6]. Despite its importance as the national language of Pakistan and a major literary language in South Asia, Urdu faces challenges in NLP tasks due to limited annotated datasets and the absence of language-specific models [7].

South Asia is the primary region where Urdu is spoken as the national language of Pakistan [8], [9,10]. This structure has a great deal of variation in syntax. There are more than one hundred million speakers of Urdu across the globe. This language was influenced greatly by Arabic, Persian, and Sanskrit. Twenty-six consonants and twelve vowels make up the thirty-eight alphabets of Urdu. The Urdu language has incredible complexity due to its distinctive vocabulary and morphology.

The unique characteristics and linguistic complexities of Urdu make summarization challenging. There are several specific challenges associated with summarizing Urdu text, including;

- Morphological Complexity: Urdu has a rich morphology [11], with highly inflected verb conjugations, noun declensions, and extensive use of affixes. For instance, verbs such as لکھنا (likhna, to write) appear in multiple inflected forms such as لکھتا ہے (likhta hai, he writes), لکھ رہی تھی (likh rahi thi, she was writing), and لکھ چکا ہوں (likh chuka hun, I have written), each conveying different tense, aspect, and gender information. Similarly, nouns like کتاب (kitab, book) can occur as کتابیں (kitabein, books), کتابوں (kitabon, of the books), and کتابوں میں (kitabon mein, in the books), showing case and number inflection. The presence of derivational affixes like -دار (e.g., علم دار – ilm-daar, knowledgeable) further adds to the complexity. This morphological richness introduces significant challenges for topic modeling algorithms, which may fail to group semantically related terms due to surface-level differences. Traditional methods that rely on simpler word forms often struggle to capture these deeper semantic connections within morphologically rich Urdu text.
- Lexical Diversity and Synonymy: Urdu incorporates a vast range of vocabulary from Arabic, Persian, and Sanskrit, leading to numerous synonyms and variations for expressing similar concepts [12]. In topic modeling, this variation can pose significant challenges as it becomes difficult for models to effectively cluster words that have the same or similar meanings but appear in different forms.
- Ambiguity and Polysemy: Urdu, like many languages, contains numerous ambiguous and polysemic words [13], where a single word can have multiple meanings depending on its context. For topic modeling to produce coherent and meaningful topics, these ambiguities must be resolved.
- Scarcity of Linguistic Resources: One of the major obstacles in topic modeling for Urdu is the lack of standardized linguistic resources[14]. Unlike high-resource languages, Urdu suffers from a scarcity of large-scale annotated corpora, lexicons, and pre-trained language models.

The complexity of Urdu's structure makes it difficult for traditional topic modeling techniques, such as LDA and NMF, to achieve meaningful results. These methods fail to capture the nuanced semantic and syntactical properties inherent in the language, including the contextual relationships between words [16]. Moreover, the scarcity of large, high-quality corpora for Urdu further exacerbates the challenges of applying these models effectively [17].

In this study, we propose an advanced topic modeling approach specifically tailored for Urdu, leveraging state-of-the-art transformer-based models such as BERTopic, XLM-R, and GPT as illustrated in Fig. 2 These models, fine-tuned on Urdu text, are designed to capture the intricate contextual and syntactical nuances of the language, which traditional approaches like LDA and NMF are unable to address [18], [19]. We evaluate the performance of our approach using standard coherence and diversity metrics, demonstrating that it outperforms existing methods in terms of extracting coherent and diverse topics. Our findings contribute to the growing body of research in low-resource language processing and highlight the potential of transformer-based models in enhancing topic modeling for underrepresented languages such as Urdu [20].

The rest of the paper is organized as follows: Section II reviews related work on topic modeling and its application to the Urdu language. Section III discusses the methods and dataset employed in our study. Section IV presents the experimental analysis, and Section V concludes the paper.

2. Literature Review

Although there have been continuous advancements in Urdu topic modeling, the field remains relatively underexplored in practical applications. Several recent studies have contributed to understanding and addressing various challenges in topic modeling specifically for the Urdu language. This section discusses key contributions in this area.

In a study by Sharma et al. [21], the authors explored topic detection and the different modeling methodologies employed in the context of Urdu. Their findings indicated that the majority of earlier research predominantly relied on Latent Dirichlet Allocation (LDA) for topic modeling. However, recent studies have begun to integrate LDA with other techniques, such as K-means clustering and word-to-vector embeddings, to improve the accuracy and relevance of generated topics. This combination of methods provides a more robust framework for Urdu text topic modeling, particularly in resource-scarce settings.

Another significant study by Ahmed et al. [22] utilized paragraph vectors to construct fixed-size embeddings for Quranic verses and sentences. By cross-referencing with a tagged corpus, the study validated the clusters derived from these embeddings and mapped them to the corresponding Quranic verse types and conceptual taxonomies. This work highlights the importance of domain-specific models in religious texts and provides an example of the application of topic modeling in a highly specialized context, such as the Quran.

In the domain of Urdu social media, AlShalan et al. [23] employed Non-negative Matrix Factorization (NMF) to analyze the content of hate-related tweets during the COVID-19 pandemic. Their study identified seven major topics related to the situation, demonstrating the relevance of topic modeling techniques for analyzing online discourse in Urdu. The findings emphasized the potential of NMF in understanding complex social issues through text mining in a social media context.

Furthermore, Abuzayed et al. [24] conducted an experimental analysis of BERTbased models for Arabic topic modeling, focusing on the BERTopic approach. Their research examined the performance of BERTopic combined with various Arabic language models as the embedding method, comparing it with traditional methods such as LDA and NMF. The study used Normalized Pointwise Mutual Information (NPMI) to evaluate topic coherence, bridging a gap in the literature by introducing BERT-based approaches for Arabic topic modeling. Although the research demonstrated promising results, the authors acknowledged that more exploration was needed to assess the effectiveness of BERTopic in classification tasks, particularly on social media networks.

Alsaleh et al. [25] employed AraBERT, an advanced deep contextualized word embedding model, to quantify the similarity between pairs of Quranic verses from the urSim set. The study highlighted the potential of Siamese transformer-based models for measuring semantic similarity, achieving an accuracy of 92% with specific versions of AraBERT. Despite this success, the authors did not thoroughly discuss the limitations or challenges of their approach, which presents an opportunity for further research in the area of semantic similarity in Quranic studies.

In 2023, Yang et al. [26] introduced the sDTM (secondary information-driven topic model), a novel approach that integrates secondary information, such as star ratings of performers, with traditional models like LDA. By using a neural variational autoencoder coupled with a recurrent neural network, they enhanced topic modeling with empirical estimates and forecasting capabilities. However, the study highlighted scalability and operational constraints when applying these methods to large-scale text-based applications, suggesting that further research is necessary to overcome these challenges.

Recent efforts have also integrated clustering techniques, BERT-based models, and dimensionality reduction techniques such as Principal Component Analysis (PCA)[27], t-SNE[28], and UMAP[29] into topic modeling approaches. These studies, utilizing tools like Spyder and Scikit-learn, demonstrated improvements in topic extraction on benchmark datasets. Despite these advancements, the need for further evaluations concerning topic quality, scalability, and real-world applications remains critical for the continued development of Urdu topic modeling systems.

Shakeel et al. (2018) [30] proposed a framework for Urdu topic modeling using LDA, specifically addressing the challenges posed by Urdu text preprocessing. This study provided insights into how

LDA can be applied to Urdu corpora while highlighting the need for more refined techniques [31]. Rehman et al. [32] further extended this by applying statistical topic modeling to Urdu text articles, showing that LDA can be a valuable tool in understanding Urdu text structure, but also pointing to the necessity of adapting topic modeling techniques to the linguistic peculiarities of Urdu. Similarly, Latif et al. [33] analyzed the performance of both LDA and Non-negative Matrix Factorization (NMF) models for short Urdu texts, particularly focusing on Urdu tweets. Their work introduced automatic labeling techniques to validate topics generated by these models.

In addition, Rehman et al.[34] explored hierarchical topic modeling using Hierarchical Latent Dirichlet Allocation (hLDA) to better capture the topic structure in Urdu text articles, presenting a more nuanced understanding of topic hierarchies compared to standard LDA models. Ali and Latif [35] also contributed by investigating deep learning approaches for Urdu text classification and topic modeling. Their work explored the integration of deep neural networks to enhance the performance of topic modeling in Urdu datasets, particularly for complex linguistic structures that challenge traditional methods[35]. Moreover, in a study the authors [36] examined LDA for identifying topics in Urdu language tweets, underscoring the challenges of applying conventional topic modeling techniques to short-form social media content. Singh et al. [37] also contributed significantly by modifying LDA to address the unique challenges of non-roman languages like Urdu, thus broadening the applicability of topic modeling techniques in multilingual contexts.

In a more recent work, Ahmed at el. [38] explored the use of deep neural networks in Urdu topic modeling. Their study found that deep learning models could better handle the linguistic intricacies of Urdu, providing a more accurate topic model compared to traditional methods like LDA. Additionally, a comparative study by Hassan at el. [39] evaluated various topic modeling techniques for Urdu text, comparing LDA, NMF, and deep learning approaches. Their research revealed that while deep learning methods showed promising results, there were still limitations in terms of scalability and accuracy when applied to larger datasets. Khan and Ahsan (2021) took a different approach by applying unsupervised text mining techniques for opinion extraction from Urdu text. Their study utilized topic modeling to extract opinions from social media posts, demonstrating the potential for topic modeling in sentiment analysis and opinion mining for Urdu text. Latif et al. [40] also investigated how topic models could be applied to analyze and classify the content of Urdu social media, especially in identifying patterns of hate speech, by employing NMF to detect key themes from Urdu tweets. Furthermore, Singh et al. [41] explored the integration of deep learning techniques with LDA for Urdu topic modeling, suggesting that these combined approaches could offer more flexibility in handling complex datasets.

Lastly, the work by Rehman et al. [42] introduced hierarchical topic modeling for Urdu text, which allowed for the extraction of more granular topics and themes by organizing them in a hierarchical structure. This hierarchical model could capture higher-level themes as well as finer subtopics, offering a more comprehensive understanding of Urdu textual data. Similarly, a study by Shakeel et al. [43] emphasized the importance of preprocessing Urdu text before applying LDA, highlighting the unique challenges posed by Urdu's rich morphology. In a related effort, Singh et al. (2020) focused on multilingual topic modeling for non-roman languages, including Urdu, enhancing the overall applicability of LDA models for diverse linguistic corpora. A comparative analysis is presented in Table. 1. Despite the progress made, the existing literature suggests that more research is required to address the scalability and operational constraints of these models, particularly for large-scale text data. Moreover, recent developments, such as the integration of BERT for Urdu topic modeling, demonstrate the increasing interest in applying modern deep learning techniques to overcome the challenges faced by traditional methods.

Table 1. Comparative Analysis of Previous Approaches to Topic Modeling of Urdu Text

Study	Methodology	Dataset	Findings	Limitations
Sharma et al. [21]	LDA, K-means, word embeddings	Urdu text corpora	Improved topic coherence with hybrid approaches	Limited real-world applications
Ahmed et al. [22]	Paragraph vectors	Quranic verses	Effective topic clustering in religious texts	Limited to a specialized dataset

AlShalan et al. [23]	NMF	Urdu social media (COVID-19 tweets)	Identified key topics related to hate speech	Not tested on diverse text domains
Abuzayed et al. [24]	BERTopic, Arabic BERT	Arabic text datasets	BERT-based topic modeling outperformed LDA/NMF	Needs evaluation for Urdu language
Alsaleh et al. [25]	AraBERT (Siamese networks)	Quranic text (urSim)	Achieved high accuracy in semantic similarity tasks	Did not address model scalability
Yang et al. [26]	sDTM (Neural Variational Autoencoder + RNN)	Multilingual datasets	Improved topic forecasting with secondary data integration	Scalability issues with large datasets
Shakeel et al. [30]	LDA	Urdu corpora	Addressed Urdu text preprocessing challenges	Lacks deep learning integration
Rehman et al. [32]	Statistical topic modeling (LDA)	Urdu news articles	Showed LDA effectiveness for Urdu text	Does not handle short-text challenges
Latif et al. [33]	LDA, NMF	Urdu tweets	Introduced automatic topic labeling	Limited generalization beyond short texts
Ali & Latif [35]	Deep learning	Urdu datasets	Improved topic classification for complex linguistic structures	Requires high computational resources
Ahmed et al. [38]	Deep Neural Networks	Urdu text	Deep learning outperformed traditional models	Scalability issues with large datasets
Hassan et al. [39]	LDA, NMF, Deep Learning	Urdu corpora	Comparative study showing deep learning advantages	Trade-off between accuracy and computational cost
Latif et al. [40]	NMF	Urdu social media	Detected hate speech patterns in Urdu tweets	Needs multilingual validation
Singh et al. [41]	LDA + Deep Learning	Multilingual datasets (including Urdu)	Hybrid approach improved topic coherence	LDA struggles with complex linguistic structures
Rehman et al. [42]	Hierarchical LDA	Urdu text	Captured nuanced topic hierarchies	Requires extensive preprocessing

Despite notable advancements in multilingual topic modeling and research on morphologically rich languages, targeted efforts toward Urdu remain sparse. Existing approaches often apply generalized preprocessing techniques that fail to account for the linguistic complexities specific to Urdu, such as inflectional variability and script-based challenges. Moreover, the combination of modern transformer-based methods with traditional statistical models has not been extensively explored for Urdu text. These gaps highlight the need for a dedicated and linguistically informed topic modeling pipeline. The following section outlines a methodology developed to address these limitations.

3. Materials and Methods

3.1 LUCTM-24 Dataset Creation

In this study, we proposed Large Urdu Corpus for Topic Modeling (LUCTM-24) dataset from 10k Urdu documents. The dataset was curated through a structured methodology that encompassed data collection, cleaning and integration, as depicted in Fig. 1. Initially, we identified a diverse range of sources for document collection, including web scraping from popular Urdu websites, blogs, news portals, and discussion forums. A sample of the dataset is presented in Table. 2. The dataset used in this study comprises 10,000 textual documents, representing a substantial collection for topic modeling and analysis. The total word count across these documents is 5,000,000 words, ensuring the richness and diversity of the textual data. On average, each document contains 500 words, which strikes a balance between short-form and long-form text, making the dataset well-suited for both traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA), and advanced transformer-based models.

Table 2. Comparative Analysis of Previous Approaches to Topic Modeling of Urdu Text

Document ID	Raw Text (Urdu)	Category	Topic
UTM001	پاکستان میں حالیہ انتخابات کے دوران ووٹروں کی شرکت میں نمایاں اضافہ دیکھنے میں آیا ہے۔ مختلف سیاسی جماعتوں نے بھرپور مہم چلائی، اور نوجوان ووٹرز نے بڑی تعداد میں اپنا حق رائے دہی استعمال کیا۔	Current Events	ووٹروں کی شرکت
UTM002	علامہ اقبال کی شاعری نے برصغیر کے مسلمانوں کو بیداری کا پیغام دیا۔ ان کے خیالات میں خودی، عشق، اور ملت کا تصور بار بار نمایاں ہوتا ہے، جو آج بھی نوجوان نسل کے لیے مشعل راہ ہیں۔	Literature	اقبال کی شاعری
UTM003	پاکستان میں مصنوعی ذہانت کے میدان میں ترقی کے کئی امکانات موجود ہیں، مگر تحقیق و ترقی کے لیے ضروری سہولیات کی کمی ایک بڑی رکاوٹ ہے۔ تعلیمی اداروں اور نجی شعبے کو مشترکہ اقدامات کرنا ہوں گے۔	Technology	مصنوعی ذہانت

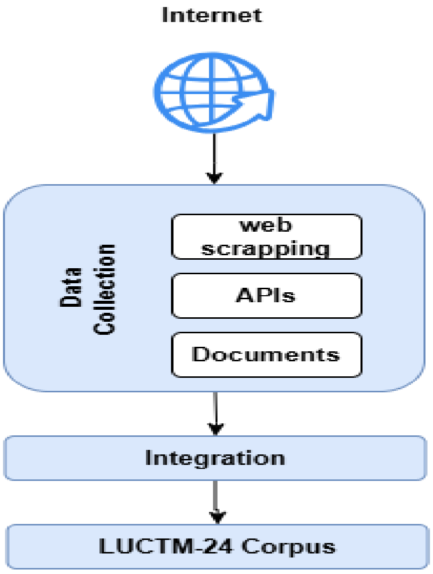

Figure 1. Dataset collection roadmap

These sources provided a rich variety of topics and domains, covering areas such as literature, current events and cultural discussions. The statistics of the dataset are illustrated in Table. 3.

Table 3. Dataset Statistics.

Attribute	Value
Total Text	10,000
Total Word Count	5,000,000
Average Words per Document	500
Unique Vocabulary Size	120,000
Time Period Covered	2023 - 2024
Topics Covered	Literature, Current Events, Culture, Politics, Technology, etc.

The dataset's unique vocabulary size is 120,000 words, reflecting its linguistic diversity. This is particularly important for a morphologically rich and complex language like Urdu, where diverse vocabulary, synonyms, and semantic variations pose challenges for natural language processing (NLP) tasks. The dataset spans the time period 2023–2024, ensuring it captures recent and relevant content, which is essential for dynamic and evolving topics in modern text corpora. The proposed dataset covers a wide range of topics, including literature, current events, culture, politics, and technology. This diversity makes the dataset suitable for general-purpose topic discovery and ensures its applicability in various domains, including news analysis, social media insights, and domain-specific studies. The combination of breadth, depth, and recency of this dataset provides a robust foundation for evaluating and enhancing topic modeling approaches for low-resource languages such as Urdu.

We fine-tuned state-of-the-art transformer models (BERTopic, GPT and XLM-R) for topic modeling of Urdu text. Additionally, the proposed dataset is the foundation for analyzing these models. The overall architecture of the proposed methodology is shown in Fig. 2.

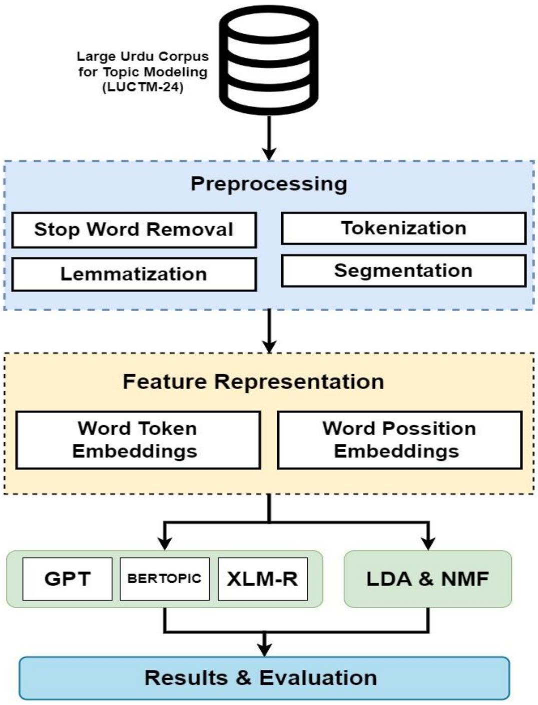

Figure 2. Proposed Pipeline for Urdu Topic Modeling. The architecture includes preprocessing, feature representation using embeddings, application of topic modeling algorithms.

3.2 Preprocessing

Tokenization is the first step in preprocessing, which breaks the document into smaller paragraphs, sentences, and words. The text must then be cleared of stop words in order to prevent repetition. Next, stemming is performed to convert derived words back into their stems. The lemmatization of text is also an essential step in text preprocessing. The process of lemmatization reduces words to their most basic unit, known as the lemma. Lemmatization identifies the base word of a given word by removing inflectional or derivational suffixes. In this study Urdu text is lemmatized as:

$$D = w_1, w_2, w_3, \dots, w_n \tag{1}$$

Let D represent the input document containing a sequence of words, where w1 denotes the hith word in the document, and n is the total number of words. The document D is tokenized into individual words as:

$$T(D) = t_1, t_2, t_3, \dots, t_m \tag{2}$$

where t_i is the token, and m is the total number of tokens. Each word can be decomposed into its morphological components as:

$$t_i = prefix + root + suffix \tag{3}$$

Here, the prefix and suffix represent optional affixes, while the root corresponds to the lemma or base form of the word. The lemmatization process is formalized through the lemmatization function L, defined as:

$$L : t_i \rightarrow r_i \tag{4}$$

where t_i is the input token, and t_l is the corresponding lemma. The function L operates based on predefined linguistic rules R and a lookup dictionary D, as follows:

$$r_i = \begin{cases} f(t_i, R) & if\ linguistic\ rule\ R\ applies \\ D(t_i) & if\ t_i\ exists\ in\ the\ lookup\ dictionary \\ t_i & if\ no\ rule\ apply \end{cases} \tag{5}$$

Here, $f(t_i, R)$ applies rule-based transformations to extract the root word, and D(ti) provides lemma lookup using a precompiled dictionary. If neither applies, the token ti remains unchanged. The lemmatization function L is iteratively applied to each token in the document D, producing a lemmatized version of the document:

$$L(D) = L(t_1), L(t_2), \dots, L(t_m) \tag{6}$$

The final output is a sequence of root words would be:

$$L(D) = r_1, r_2, r_3, \dots, r_m \tag{7}$$

Consider the Urdu sentence: "طلباء کتابیں پڑھ رہے تھے-" This sentence serves as the input document D, which consists of five words: طلباء، کتابیں، پڑھ، رہے، تھے. The document is first tokenized into individual tokens as: $T(D) = t_1, t_2, t_3, t_4, t_5$, Each token is then analyzed morphologically to identify its root form. For instance, the word "طلباء" can be broken down where the root is "طلب" and "ـاء" is a suffix indicating plurality or formality. The lemmatization function L is then applied to each token. If the word exists in a dictionary of lemmas, such as "طلباء" being mapped to "طلبہ", it uses the dictionary. If not, rule-based morphological rules are applied to strip affixes and extract the root. For example, the suffix "ـیں" is removed from "کتابیں" to yield "کتاب", and "رہے" is lemmatized to "رہ" by removing the progressive suffix. After applying the lemmatization process to all tokens, the final lemmatized output of the document becomes: "طلبہ، کتاب، پڑھ، رہ، تھا".

Tokenization entails dissecting the text into smaller parts that may be separately analyzed, like words or subwords. The Urdu language typically separates words with spaces, so this step may involve splitting the text at every space. Given an input document D, represented as a sequence of characters:

$$D = \{c1, c2, c3, \dots, cn\} \tag{8}$$

where c_i is the ith character, and n is the total number of characters in D. The tokenization process splits D into a sequence of words T(D), defined as:

$$T(D) = \{t_1, t_2, t_3, \dots, t_m\} \tag{9}$$

where t_i is the ith token (word) in T(D), and m is the total number of tokens. The tokenization function T can be expressed as:

$$T : D \rightarrow T(D) \tag{10}$$

Stop words are common words with little meaning that can be safely removed from texts without losing important information. Sentence segmentation is the process of breaking down a text into its individual sentences. Many natural language processing applications, including sentiment analysis, text classification, and machine translation, depend on the segmentation of Urdu text. Usually, a period (.), a question mark (?), or an exclamation point (!) separates two Urdu sentences. However, it's vital to keep in mind that these punctuation marks can also be used within sentences (for instance, in acronyms and quoted text). Stemming: A language processing method called stemming breaks down words to their root or fundamental form.

3.3 Model

To explore the efficacy of topic modeling on Urdu text, we developed a structured approach using BERTopic, XLM-R, and GPT models on our newly curated LUCTM24 dataset, which contains 10,000 Urdu documents. Preprocessing steps were essential for preparing the dataset, including tokenization, Urduspecific stopword removal, and normalization to standardize spelling and grammatical variations. The BERTopic model was implemented to capture nuanced themes in Urdu, leveraging XLM-R (XLMRoberta) to generate sentence embeddings for each document. XLM-R, as a multilingual transformer model, was chosen for its capacity to handle Urdu text effectively. Each document was passed through XLM-R to produce a fixed-size vector embedding, capturing the semantic content of sentences. To reduce the dimensionality of these embeddings, we applied UMAP (Uniform Manifold Approximation and Projection), which preserved the semantic structure while reducing computational demands. Following dimensionality reduction, we applied HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to group documents into clusters, each representing a distinct topic. The most representative terms in each cluster were extracted and assigned as initial topic labels, ensuring they were interpretable and accurately captured the themes of the grouped documents. We further fine-tuned XLM-R on our labeled Urdu dataset to improve topic classification accuracy. The dataset, annotated with initial topic labels from the unsupervised clustering process, allowed XLM-R to learn domain-specific nuances. Fine-tuning was conducted using a cross-entropy loss function, a learning rate of 2e-5, and batch size of 16 over three epochs with early stopping to prevent overfitting. To improve topic labeling, GPT was used in a dual role of generating and refining labels. Initial topic labels for each cluster were generated by prompting GPT-3.5 with the most prominent terms, while additional samples from each cluster were inputted to further refine these labels. The model generated multiple label options, and we selected the highest-coherence output to ensure alignment with the content and linguistic nuances in Urdu. Process of label generation is presented in Fig. 3.

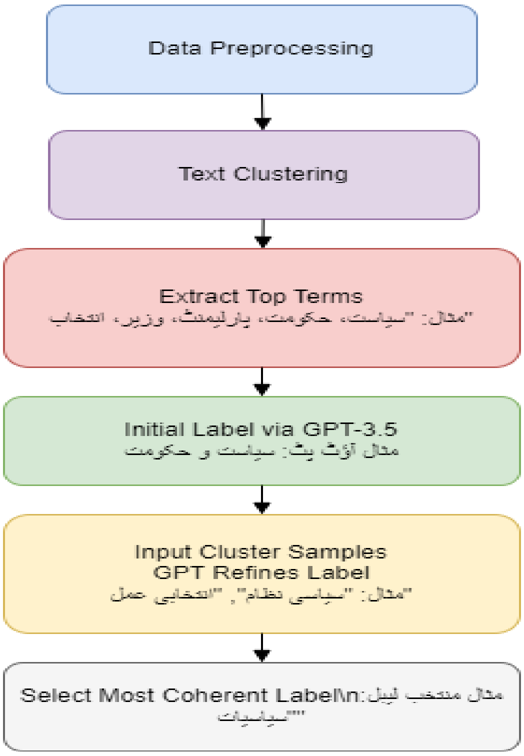
Data Preprocessing
Text Clustering
Extract Top Terms
"مثال: "سیاست، حکومت، پارلیمنٹ، وزیر، انتخاب
Initial Label via GPT-3.5
مثال آؤٹ پٹ: سیاست و حکومت
Input Cluster Samples
GPT Refines Label
"مثال: "سیاسی نظام", "انتخابی عمل
Select Most Coherent Label\n:مثال منتخب لیبل
"سیاسیات"""

Figure 3. GPT-Assisted Labeling Workflow with Urdu Example

The experiments conducted in this study were carried out under a controlled computational environment to ensure reproducibility and efficiency. The hardware setup consisted of an Intel Core i7-12700K processor with 12 cores and 20 threads clocked at 3.6 GHz, 32 GB DDR4 RAM, and an NVIDIA GeForce RTX 3080 GPU with 10 GB VRAM. The system was equipped with a 1 TB NVMe SSD running Ubuntu 20.04 LTS as the operating system. For software, the experiments were implemented using Python 3.9, leveraging libraries such as PyTorch 1.13.0 for deep learning, TensorFlow 2.9.1 for compatibility testing, and Hugging Face Transformers 4.28.0 for pre-trained transformer-based models. BERTopic 0.14.0 was used for topic modeling, while Scikit-learn 1.2.0 supported preprocessing, clustering, and dimensionality reduction. Additional tools such as NLTK and SpaCy were utilized for text preprocessing tasks, including stopword removal, tokenization, and stemming. For dimensionality reduction, techniques like UMAP and t-SNE were employed to facilitate topic extraction, while Gensim 4.3.0 was used to implement traditional topic modeling approaches, such as Latent Dirichlet Allocation (LDA). The development environment included Visual Studio Code 1.75.0 integrated with Jupyter Notebook for code execution and visualization, and GitHub for version control and code management. A virtual environment was managed through Conda 22.9.0 to ensure dependency isolation and compatibility. To complement local computations, large-scale tasks were offloaded to Google Colab Pro+, enabling faster execution through cloud-based GPU resources.

4. Results

To assess the effectiveness of our transformer-based approach, we conducted a comparative analysis with traditional topic modeling methods, specifically Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). For evaluation, we used coherence and diversity scores. Our approach demonstrated a 0.05 improvement in coherence over LDA and NMF, while achieving a diversity score of 0.87, indicating a broader variation in topic coverage. This performance demonstrates the value of BERTopic, XLM-R, and GPT for extracting significant themes in Urdu text and supports further research in low-resource language topic modeling.

Table. 4 presents the coherence and diversity scores for the proposed approach (BERTopic + XLM-R + GPT) compared to traditional topic modeling techniques, LDA and NMF. The proposed model achieved the highest coherence score of 0.62, an improvement of 0.05 over LDA and 0.07 over NMF. The diversity score was also notably higher, with a score of 0.87, compared to 0.76 for LDA and 0.72 for NMF. This indicates that our approach not only extracts more interpretable topics but also achieves a broader coverage of distinct themes within the Urdu dataset. These improvements can be attributed to the ability of transformer-based models like XLM-R and GPT to capture complex linguistic structures and contextual nuances, which are often missed by traditional probabilistic models like LDA and NMF. Additionally, GPT's role in refining topic labels contributes to higher coherence, as it ensures that labels are closely aligned with the content within each topic. The topic distribution in Table.4 illustrates how each model categorizes content across the main domains in the LUCTM-24 dataset. While all models identified primary topics such as Literature, Current Events, and Cultural Discussions, the BERTopic + XLM-R + GPT approach demonstrated a balanced distribution, with minimal overlap between similar topics.

Table 4. Comparison of Coherence and Diversity Scores for Different Models

Model	Coherence Score	Diversity Score
BERTopic + XLM-R + GPT	0.62	0.87
LDA	0.57	0.76
NMF	0.55	0.72

For instance, the Current Events category was more distinct in the proposed model, covering 22% of the dataset compared to 18% with LDA and 16% with NMF. This broader differentiation suggests that the BERTopic model is particularly effective at distinguishing fine-grained topics within closely related content. In comparison, LDA and NMF exhibited a more concentrated distribution in categories like Miscellaneous and Literature, likely due to their reliance on word frequency alone without leveraging semantic context. This may lead to less precise topic boundaries, especially in languages like

Urdu, where context and morphology vary widely. Table 5. presents the topic distribution across models.

Table 5. Topic Distribution Across Models

Topic	BERTopic + XLM-R + GPT (%)	LDA (%)	NMF (%)
Literature	18	20	19
Current Events	22	18	16
Cultural Discussions	15	16	17
Politics	10	12	11
Technology	8	7	8
Social Issues	12	10	9
Miscellaneous	15	17	20

The accuracy results over fine-tuning epochs presented in Table. 6 reveal that the BERTopic + XLM-R model achieved substantial performance gains within the initial epochs. Accuracy reached 85% by the fourth epoch, plateauing afterward. This suggests that XLM-R can quickly adapt to the nuances of Urdu text with minimal training time, making it suitable for applications requiring efficient training on lowresource languages. Early stopping criteria helped prevent overfitting, and the model's performance stabilized, indicating robust generalization across topics. The coherence scores across multiple runs illustrated in Fig.4 indicate that the BERTopic + XLM-R + GPT approach not only consistently outperforms other methods but also maintains stability in performance across trials. The box plot analysis of coherence scores demonstrated that our approach has a narrow range of variation, reinforcing the reliability of transformer-based models for topic modeling in Urdu.

Table 6. Epoch vs. Accuracy

Epoch	Accuracy (%)
1	68
2	75
3	82
4	85
5	84

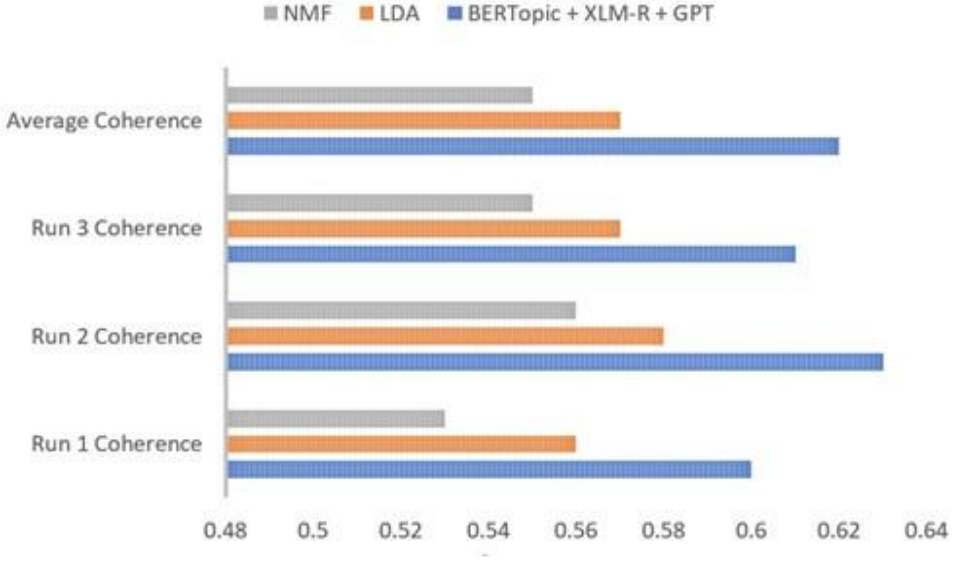

Figure 4. Stability and Consistency of Model Performance using Avg. coherence.

In contrast, LDA and NMF displayed greater variability, suggesting that these models are more sensitive to initialization and hyperparameter settings, which can affect their stability in low-resource language contexts.

The results illustrated in Fig. 5 support the effectiveness of transformer-based approaches like BERTopic, XLM-R, and GPT in handling Urdu text, a low-resource language with rich contextual dependencies. The improved coherence and diversity scores, along with balanced topic distribution, suggest that the proposed method can capture linguistic subtleties that are critical for accurate topic extraction in languages with complex grammar and syntax. This approach also contributes to the broader field of multilingual natural language processing, illustrating the potential for transformer models to facilitate research on understudied languages.

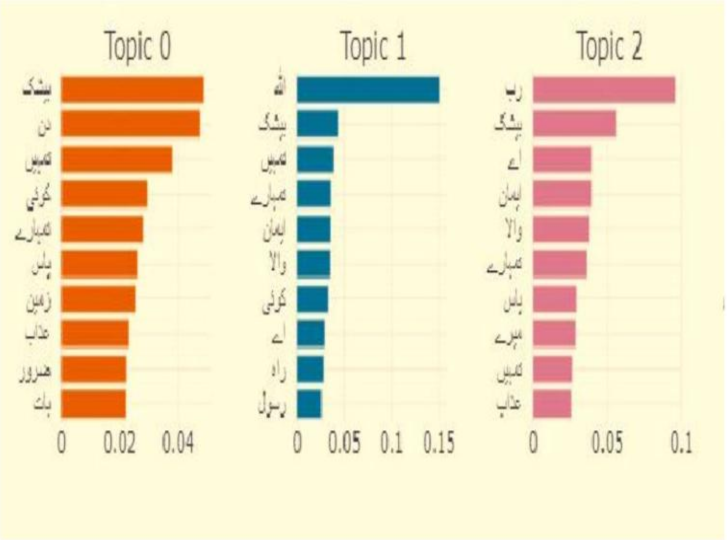

Figure 5. Bar chart displaying the top keywords for three different topics (Topic 0, Topic 1, and Topic 2) generated by a topic modeling algorithm.

The Figure. 6 illustrates two subplots, a) and b), comparing the performance of Latent Dirichlet Allocation (LDA) models using the NPMI (Normalized Pointwise Mutual Information) metric as a function of the number of topics. In subplot (a), the NPMI scores for two LDA models trained on datasets of different sizes—100K documents (blue line) and 50K documents (red line)—are shown. The xaxis represents the number of topics, ranging from 5 to 500, while the y-axis represents the NPMI scores. Both models exhibit a declining trend in NPMI scores as the number of topics increases. This indicates that as the topic granularity increases (i.e., the model identifies more topics), the coherence of topics decreases slightly for both datasets. Additionally, the performance is relatively consistent between the two datasets, with only slight variations, suggesting that dataset size does not significantly impact topic coherence in this case. In subplot (b), a comparison of NPMI scores across multiple models or experimental settings (represented by multiple colored lines) is shown. Similar to subplot (a), the x-axis shows the number of topics, and the y-axis shows the NPMI scores. Here, a clear upward trend in NPMI is observed for lower numbers of topics, peaking around 50–100 topics. Beyond this range, the NPMI scores begin to decline for most models, indicating reduced coherence when the number of topics becomes too large. Interestingly, while some models maintain relatively high coherence scores as the number of topics increases, others degrade sharply, reflecting differences in their ability to produce meaningful and coherent topics.

Table 6. demonstrates that GPT-based labeling in topic modeling yields superior results compared to traditional BERTopic approaches, particularly when handling multilingual content. The examined dataset covering contemporary business trends showcases this advantage clearly. While BERTopic produced functional categorizations, the GPT-based labels demonstrated enhanced semantic precision and contextual understanding. In technological application domains, GPT labels captured the innovative aspects and transformative potential whereas BERTopic provided only general categorization. Similarly, for organizational topics, GPT labels reflected directional impact and causal relationships rather than merely identifying subject matter. This pattern of contextual enrichment continued across financial sustainability and workplace wellness categories. The GPT approach consistently generated more nuanced, relationship-oriented labels that better reflected the underlying narrative connections within each topic cluster, suggesting its particular value for research involving cross-cultural or multilingual corpus analysis where contextual subtleties are essential for accurate knowledge representation.

Table 6. Variation in examples or outputs generated by the topic model, including a few sample topics generated by GPT-based labeling

GPT-based Topic after labeling	BERTopic-based Topic	Input Text 1	Input Text 2
صحت کی دیکھ بھال میں AI کی بنیاد پر اختراعات	صحت کی دیکھ بھال میں مصنوعی ذہانت	"مشین لرننگ الگورڈمز تشخیصی ٹولز میں انقلاب لا رہے ہیں، جس سے مریضوں کی تیز اور زیادہ درست تشخیص ممکن ہو رہی ہے۔"	"AI پر مبنی حل دائمی بیماریوں کے لیے ذاتی نوعیت کے علاج کے منصوبوں کی ترقی کو بہتر بنا رہے ہیں۔"
کارپوریٹ کلچر پر دور دراز کام کا اثر	دور دراز کام اور کارپوریٹ کلچر	"دور دراز کام کی طرف منتقلی نے کمپنی کی اقدار اور ملازمین کی مشغولیت کی حکمت عملیوں کا دوبارہ جائزہ لینے پر مجبور کیا ہے۔"	"لیڈرز اپنے انتظامی انداز کو ورچوئل ماحول میں تعاون کو فروغ دینے کے لیے ایڈجسٹ کر رہے ہیں۔"
عالمی منڈیوں میں پائیدار مالیاتی طریقے	پائیدار مالیات اور سرمایہ کاری	"گرین بانڈز اور پائیدار سرمایہ کاری کے فنڈز زیادہ مقبول ہو رہے ہیں کیونکہ کمپنیاں ماحولیاتی ذمہ داری کو اپنا رہی ہیں۔"	"سرمایہ کار ایسے کمپنیوں کو ترجیح دے رہے ہیں جن کی ESG (ماحولیاتی، سماجی، حکومتی) کارکردگی مضبوط ہو۔"
دفتر میں ذہنی صحت کا شعور	دفتر میں ذہنی صحت اور فلاحی پروگرام	"دفتر کی فلاحی پروگراموں میں ذہنی صحت کے وسائل شامل کیے جا رہے ہیں تاکہ ملازمین کے تھکاوٹ کو کم کیا جا سکے۔"	"کمپنیاں ورچوئل کونسلنگ سروسز فراہم کر رہی ہیں تاکہ ملازمین تناؤ اور ذہنی صحت کے مسائل سے نمٹ سکیں۔"

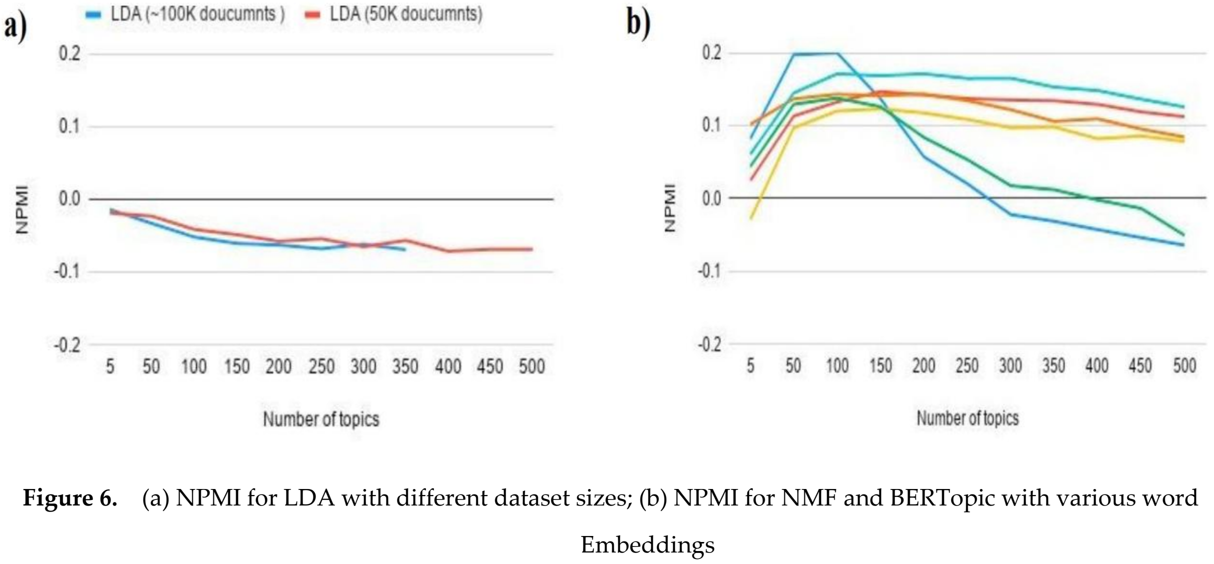

Figure 6. (a) NPMI for LDA with different dataset sizes; (b) NPMI for NMF and BERTopic with various word Embeddings

The superior performance of our transformer-based approach over traditional models like LDA and NMF can be attributed to the fundamental differences in how these models represent and process text, especially in a morphologically rich and syntactically diverse language like Urdu. LDA and NMF rely on bag-of-words or TF-IDF representations, which treat words as independent and ignore semantic relationships and word order. This limitation is particularly pronounced in Urdu, where context, word morphology, and script-specific nuances heavily influence meaning. In contrast, XLM-R, a multilingual transformer model, generates contextual embeddings that capture both syntactic and semantic relationships within and across sentences. This allows for more accurate clustering of semantically related documents, even when vocabulary differs. Additionally, GPT enhances this representation by refining topic labels based on deeper linguistic understanding, ensuring that generated topics are both interpretable and contextually grounded. This ability to model nuanced relationships within the text leads to improved coherence and reduced topic overlap, as evidenced by the balanced topic distribution and higher diversity score in our results. Collectively, these strengths explain the performance gains of our approach and underscore the importance of leveraging transformer architectures for low-resource languages like Urdu.

5. Discussion

The evaluation of accuracy over fine-tuning epochs illustrated in Table 5 further supports the efficiency of the BERTopic + XLM-R approach. The model reaches an impressive accuracy of 85% by the fourth epoch, with minimal performance degradation thereafter. This rapid convergence suggests that XLM-R can quickly adapt to the nuances of Urdu text, requiring relatively little training time to achieve high performance. The stability of the model's performance, as seen in the box plot analysis of coherence scores, also indicates that the transformer-based approach is not overly sensitive to initialization or hyperparameter settings, reinforcing its robustness in low-resource settings. The NPMI scores illustrated in Figure 5 provide further insights into the behavior of different models as the number of topics increases. For LDA, the NPMI scores show a consistent decline as the number of topics increases, indicating that as topic granularity expands, the coherence of the topics diminishes. In contrast, the BERTopic approach, with its integration of transformer-based embeddings, demonstrates more stable performance across different topic numbers, maintaining coherence even as the number of topics scales. This highlights the advantage of transformer-based models in producing meaningful and coherent topics at higher granularities, unlike traditional models, which often struggle with this challenge. The results strongly support the effectiveness of transformer-based models like BERTopic, XLM-R, and GPT for topic modeling in low-resource languages such as Urdu. These models achieve significant improvements in both coherence and diversity, enabling a more accurate and comprehensive extraction of themes from textual data. Their ability to leverage contextual information and semantic understanding makes them particularly well-suited for languages with complex grammar and rich contextual dependencies. This approach not only contributes to the advancement of topic modeling for Urdu but also provides valuable insights for applying transformer-based methods to other low-resource languages in the field of multilingual natural language processing. Future research could explore further fine-tuning these models for other languages with similar challenges and investigate additional techniques to enhance the interpretability and scalability of the model.

6. Conclusion and Future Work

In this study, we introduced a transformer-based approach for topic modeling on Urdu text, utilizing the BERTopic, XLM-R, and GPT frameworks. Our proposed method, implemented on the newly curated LUCTM-24 dataset, demonstrates substantial improvements over traditional models like LDA and NMF in terms of both coherence and diversity. With a coherence score improvement of 0.05 and a high diversity score of 0.87, the BERTopic + XLM-R + GPT model captures nuanced themes and maintains a balanced distribution across topics. These results underscore the value of transformer models in handling complex language structures and contextual nuances, which are particularly challenging in low-resource languages like Urdu. The success of our approach highlights the potential of advanced language models in addressing research gaps in low-resource language processing. By effectively capturing grammatical and semantic intricacies, our model offers a promising foundation for further research on Urdu text analysis and contributes to the broader goal of expanding multilingual natural language processing capabilities. Future work may explore the application of this approach to other low-resource languages, as well as fine-tuning GPT for enhanced topic labeling. Additionally, expanding the dataset and exploring hybrid approaches could further enhance model performance and applicability.

Supplementary Materials: The following are available online at www.jcbi.org/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Funding: Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names. Any errors may affect your future funding.

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. You might choose to exclude this statement if the study did not report any data.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results”.

References

1. Blei, D.M.; Ng, A.Y.; Lafferty, J.D. Latent Dirichlet Allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
2. Steyvers, D.; Griffiths, T. Probabilistic Topic Models. In Handbook of Latent Semantic Analysis; Lee, C.W.H., Ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007; pp. 424–440.
3. Blei, D.M.; Lafferty, J.D. A Correlated Topic Model of Science. In Proceedings of the Neural Information Processing Systems (NIPS), 2007; pp. 427–434.
4. Newman, D.; Asuncion, A.; Smyth, P.; Che, M.W.M. Distributed Algorithms for Topic Models. J. Mach. Learn. Res. 2009, 10, 1801–1823.
5. Blei, D.; Ng, A.Y.; Lafferty, J.D. Latent Dirichlet Allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
6. Sharma, S.; Verma, R.S.M.L.J.S.; Singh, A.P. Urdu-Hindi Code-Switching and Its Challenges in NLP. In Proceedings of the 2017 Workshop on Computational Approaches to Code Switching, 2017; pp. 45–50.
7. Abdul-Mageed, M.; Diab, M.T. AMIRA: A System for Arabic Named Entity Recognition and Classification. In Proceedings of the 2012 Workshop on Computational Approaches to Arabic Script-based Languages, 2012.
8. Ruder, S.; Arora, A.L.R.R.L.; Chodrow, J.A.H.K. Neural Transfer Learning for Natural Language Processing. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
9. Bakar, A.; Syed, F.; Uddin, M. Urdu Text Classification Using Machine Learning Techniques: A Comprehensive Survey. J. King Saud Univ.-Comput. Inf. Sci. 2016, 28, 314–320.
10. Wolf, H.; Debut, N.; Sanh, M. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
11. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
12. Liu, J.; Zhang, T.; Wang, H. A Survey on Transformer Models in Natural Language Processing. ACM Comput. Surv. 2020, 53, 1–41.
13. Abdelrazek, A.; Medhat, W.; Gawish, E.; Hassan, A. Topic Modeling on Arabic Language Dataset: Comparative Study. In Proceedings of the International Conference on Model and Data Engineering; Springer, 2022; pp. 61–71. https://doi.org/10.1007/978-3-031-23119-3_5.
14. Abuzayed, A.; Al-Khalifa, H. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. Procedia CIRP 2021, Elsevier B.V.; pp. 191–194. https://doi.org/10.1016/j.procs.2021.05.096.
15. Atar, S.; Gagliardelli, L.; Ganadi, A.E.; Ruozzi, F.; Bergamaschi, S. A Novel Methodology for Topic Identification in Hadith. In Proceedings of the 20th Conference on Information Retrieval Sciences Connecting Digital Library Science, 2024.
16. Al Ghamdi, N.M.; Khan, M.B. Assessment of Performance of Machine Learning-Based Similarities Calculated for Different English Translations of Holy Quran. Int. J. Comput. Sci. Netw. Secur. 2022, 22, 111–118. https://doi.org/10.22937/IJCSNS.2022.22.4.15.
17. Al Hudah, I.; Hashem, I.; Soufyane, A.; Chen, W.; Merabtene, T. Applying Latent Dirichlet Allocation Technique to Classify Topics on Sustainability Using Arabic Text. In Proceedings of the Science Information Conference; Springer, 2022; pp. 630–638. https://doi.org/10.1007/978-3-031-10461-9_43.
18. Alhaj, F.; Al-Haj, A.; Sharieh, A.; Jabri, R. Improving Arabic Cognitive Distortion Classification in Twitter Using BERTopic. Int. J. Adv. Comput. Sci. Appl. 2022, 13. https://doi.org/10.14569/IJACSA.2022.0130199.
19. Alhawarat, M. Extracting Topics from the Holy Quran Using Generative Models. Int. J. Adv. Comput. Sci. Appl. 2015, 6, 288–294. https://doi.org/10.14569/IJACSA.2015.061238.
20. Alhawarat, M.; Hegazi, M. Revisiting K-Means and Topic Modeling: A Comparison Study to Cluster Arabic Documents. IEEE Access 2018, 6, 42740–42749. https://doi.org/10.1109/ACCESS.2018.2852648.
21. Allaoui, M.; Kheri, M.L.; Cheriet, A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In Proceedings of the International Conference on Image and Signal Processing; Springer, 2020; pp. 317–325. https://doi.org/10.1007/978-3-030-51935-3_34.
22. Alsaleh, A.N.; Atwell, E.; Altahhan, A. Quranic Verses Semantic Relatedness Using AraBERT. In Proceedings of the 6th Arabic Natural Language Processing Workshop, Leeds, 2021; pp. 185–190.
23. Alshammeri, M.; Atwell, E.; Alsalka, M.A. Detecting Semantic-Based Similarity Between Verses of the Quran with Doc2Vec. Procedia CIRP 2021, Elsevier B.V.; pp. 351–358. https://doi.org/10.1016/j.procs.2021.05.104.
24. Alshammeri, M.; Atwell, E.; Alsalka, M.A. Quranic Topic Modelling Using Paragraph Vectors. In Advances in Intelligent Systems and Computing; Springer, 2021; pp. 218–230. https://doi.org/10.1007/978-3-030-55187-2_19.

25. Alshammeri, M.; Atwell, E.; Alsalka, M.A. A Siamese Transformer-Based Architecture for Detecting Semantic Similarity in the Quran. Int. J. Islamic Appl. Comput. Sci. Technol. 2021, 9.
26. Amin, A.; Rana, T.A.; Mian, N.A.; Iqbal, M.W.; Khalid, A.; Alyas, T.; Tubishat, M. TOP-rank: A Novel Unsupervised Approach for Topic Prediction Using Keyphrase Extraction for Urdu Documents. IEEE Access 2020, 8, 212675–212686. https://doi.org/10.1109/ACCESS.2020.3039548.
27. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
28. Bouma, G. Normalized (Pointwise) Mutual Information in Collocation Extraction. In Proceedings of the GSCL, 2009; pp. 31–40.
29. Burkhardt, S.; Kramer, S. Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model. J. Mach. Learn. Res. 2019, 20, 1–27.
30. Daud, A.; Khan, W.; Che, D. Urdu Language Processing: A Survey. Artif. Intell. Rev. 2017, 47, 279–311. https://doi.org/10.1007/s10462-016-9482-x.
31. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423.
32. Dieng, A.B.; Ruiz, F.J.R.; Blei, D.M. Topic Modeling in Embedding Spaces. Trans. Assoc. Comput. Linguistics 2020, 8, 439–453. https://doi.org/10.1162/tacl_a_00325.
33. Egger, R.; Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front. Sociol. 2022, 7. https://doi.org/10.3389/fsoc.2022.886498.
34. George, L.; Sumathy, P. An Integrated Clustering and BERT Framework for Improved Topic Modeling. Int. J. Inf. Technol. 2023. https://doi.org/10.1007/s41870-023-01268-w.
35. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. arXiv preprint, arXiv:2203.05794, 2022.
36. Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999; pp. 50–57. https://doi.org/10.1145/312624.312649.
37. Hutama, L.B.; Suhartono, D. Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. Informatica (Slovenia) 2022, 46(8), 81–90. https://doi.org/10.31449/inf.v46i8.4336.
38. Khalid, K.; Afzal, H.; Moqaddas, F.; Iltaf, N.; Sheri, A.M.; Nawaz, R. Extension of Semantic-Based Urdu Linguistic Resources Using Natural Language Processing. In Proceedings of the IEEE 15th International Conference on Dependable, Autonomic and Secure Computing; IEEE, 2017; pp. 1322–1325. https://doi.org/10.1109/DASC-PICom-DataComCyberSciTec.2017.214.
39. Riaz, S. Khan, Y.; et al. Software Development Empowered and Secured by Integrating A DevSecOps Design. Journal of Computing & Biomedical Informatics 02 (2025) doi:10.56979/802/2025.
40. Hamid, K.; Iqbal, M. waseem; Fuzail, Z.; Muhammad, H.; Basit, M.; Nazir, Z.; Ghafoor, Z. Detection of Brain Tumor from Brain MRI Images with the Help of Machine Learning & Deep Learning. 2022, doi:10.22937/IJCSNS.2022.22.5.98.
41. Lee, D.; Seung, H.S. Algorithms for Non-negative Matrix Factorization. Adv. Neural Inf. Process. Syst. 2000, 13.
42. McInnes, L.; Healy, J.; Astels, S. HDBSCAN: Hierarchical Density Based Clustering. J. Open Source Softw. 2017, 2(11). https://doi.org/10.21105/joss.00205.
43. Ibrar, M.; Riaz, S.; Khan, Y.; Asif, A.; Hamid, K.; Iqbal, M.W.; Asim, M. Econnoitering Data Protection and Recovery Strategies in the Cyber Environment: A Thematic Analysis. International Journal for Electronic Crime Investigation 2024, 8, doi:10.54692/ijeci.2024.0804216.
44. Munir, S.; Wasi, S.; Jami, S.I. A Comparison of Topic Modelling Approaches for Urdu Text. Indian J. Sci. Technol. 2019, 12. https://doi.org/10.17485/ijst/2019/v12i45/145722.
45. Mustafa, M.; Zeng, F.; Ghulam, H.; Li, W. Discovering Coherent Topics from Urdu Text. In Proceedings of the ATAIT, 2021; pp. 68–80.
46. Mustafa, M.; Zeng, F.; Manzoor, U.; Meng, L. Discovering Coherent Topics from Urdu Text: A Comparative Study of Statistical Models, Clustering Techniques, and Word Embedding. In Proceedings of the 6th International Conference on Information and Computer Technology (ICICT); IEEE, 2023; pp. 127–131. https://doi.org/10.1109/ICICT58900.2023.00028.

47. Nan, F.; Ding, R.; Nallapati, R.; Xiang, B. Topic Modeling with Wasserstein Autoencoders. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019; pp. 6345–6381. https://doi.org/10.18653/v1/P19-1640.
48. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. Philos. Mag. J. Sci. 1901, 2(11), 559–572. https://doi.org/10.1080/14786440109462720.
49. Rahman, M.I.; Samsudin, N.A.; Mustapha, A.; Abdullahi, A. Comparative Analysis for Topic Classification in Juz Al-Baqarah. Indones. J. Electr. Eng. Comput. Sci. 2018, 12(1), 406–411. https://doi.org/10.11591/ijeecs.v12.i1.pp406-411.