

Exploring topic modelling: a comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity

Ayesha Riaz¹, Omar Abdulkader², Muhammad Jawad Ikram², Sadaqat Jan¹

¹Department of Computer Software Engineering, University of Engineering and Technology, Mardan, Pakistan

²Faculty of Computer Studies, Arab Open University, Jeddah, Saudi Arabia

Article Info

Article history:

Received May 12, 2024

Revised Oct 30, 2024

Accepted Nov 20, 2024

Keywords:

Bidirectional encoder representations from transformers

Extra long-term memory networks

Generative pre-trained transformers

Latent Dirichlet allocation

Social media platforms

ABSTRACT

Topic modeling (TM) is an unsupervised technique used to recognize hidden or abstract topics in large corpora, extracting meaningful patterns of words (semantics). This paper explores TM within data mining (DM), focusing on challenges and advancements in extracting insights from datasets, especially from social media platforms (SMPs). Traditional techniques like latent Dirichlet allocation (LDA), alongside newer methodologies such as bidirectional encoder representations from transformers (BERT), generative pre-trained transformers (GPT), and extra long-term memory networks (XLNet) are examined. This paper highlights the limitations of LDA, prompting the adoption of embedding-based models like BERT and GPT, rooted in transformer architecture, offering enhanced context-awareness and semantic understanding. The paper emphasizes leveraging pre-trained transformer-based language models to generate document embedding, refining TM and improving accuracy. Notably, integrating BERT with XLNet summaries emerges as a promising approach. By synthesizing insights, the paper aims to inform researchers on optimizing TM techniques, potentially shifting how insights are extracted from textual data.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Omar Abdulkader

Faculty of Computer Studies, Arab Open University

Jeddah, Saudi Arabia

Email: o.abdulkader@arabou.edu.sa

1. INTRODUCTION

Artificial intelligence (AI) is a comprehensive field in computer science that encompasses various technologies and methodologies to enable machines to perform tasks that traditionally required human intelligence. Machine learning (ML) is a subset of AI, focusing on the development of algorithms that allow machines to learn patterns and make predictions or decisions from data without explicit programming. ML enhances performance through autonomous learning. On the other hand, natural language processing (NLP), another crucial component of AI, involves the application of techniques to enable machines to understand and process human language in spoken or written form. NLP facilitates speech-based interactions between humans and computers, addressing the gap between human communication and computer comprehension. It finds applications in sentiment analysis, text classification, semantic relatedness, or topic modeling. Due to the rapid growth of online textual information such as web pages, email, news articles, and personal blogs, there is a need to search, analyze, and understand the massive amount of data to determine the subjective information in text. Because of the tremendous amount of data on online platforms, it is a difficult and time-consuming task for individuals to extract information about the topic of their interest. Different machine

learning algorithms such as text summarization and topic modeling approaches are used for analysis of such data. To analyze documents to learn meaningful patterns of words, topic modeling is used. Extracting topics from text is an important application. Topic modeling is a method used for finding the unique topics that are present in a dataset. It is a text-mining application [1]. As depicted in Figure 1, the process of topic modeling initiates with the acquisition of raw input text data. This data then proceeds through a crucial data preprocessing phase aimed at cleaning and organizing it for subsequent analysis. Once the preprocessing stage is complete, various topic modeling techniques are employed to extract coherent and meaningful topics from the refined text corpus. Subsequently, these derived topics are visualized, facilitating exploration and interpretation and offering valuable insights into the inherent structures and themes within the text. The probabilistic topic models are used to discover the secret thematic structure in a corpus of documents and help in recognizing the massive amount of unlabeled data [2]. This technique is used to get word co-occurrence patterns across a collection of documents; these patterns are then considered as hidden “topics” present in the corpus.

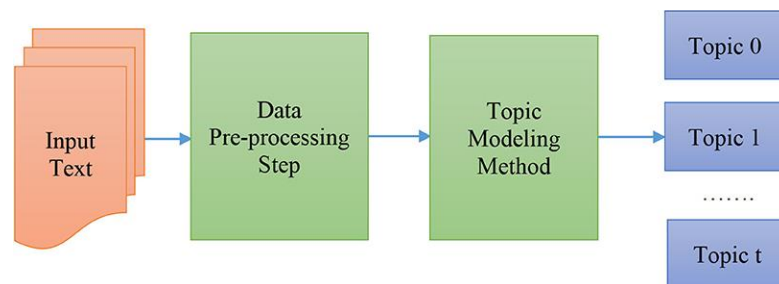


Figure 1. Topic modeling

Topic models, like latent semantic analysis (LSA) which utilizes dimensionality reduction for uncovering latent structures in text corpora based on term-document relationships. On the other hand, latent Dirichlet allocation (LDA) is a probabilistic model representing documents as mixtures of topics. While it incorporates probabilistic modeling. Traditional topic models, such as latent Dirichlet allocation [3] and probabilistic latent semantic analysis [3], have shown to be an effective unsupervised method for the statistical analysis of document collections. These approaches [5], [6] model each document as a mixture of latent topics, which are multinomial distributions over words, and adhere to the bag-of-words assumption. Non-negative matrix factorization (NMF) an alternative to probabilistic models, decomposing documents into topics with non-negative coefficients.

A novel approach to unsupervised word embedding has recently been introduced, offering a unique representation for individual word types as vectors. This method facilitates the clustering of words based on their distances within a high-dimensional space. By providing a vector representation for each word type, this approach enhances the understanding of semantic relationships between words and enables more effective analysis and interpretation of textual data. A popular word representation technique that can be used to identify themes with semantic meaning is word embedding. It produces word vector representations, and research has demonstrated that it is an effective indicator of semantic relatedness [7].

Word embedding [8], a technology for natural language processing that is now undergoing rapid development, allow us to model themes and topic connections in continuous semantic space. Word embedding are real-valued continuous vectors for words that are successful at capturing semantic regularities in language. They are also referred to as word vectors and distributed representations of words. Similar semantic and syntactic characteristics of words tend to project into surrounding regions in the vector space. Embedding-based models in topic modeling, such as Doc2Vec, represent a paradigm shift from traditional probabilistic approaches. Doc2Vec introduced in 2014, extends Word2Vec, enabling the representation of entire documents in continuous vector spaces.

Over time, these models have evolved with refinements in training methodologies, incorporation of additional contextual information, and optimization for performance and scalability. Researchers have explored advancements in embedding techniques, enhancing the modeling of semantic relationships and capturing more nuanced document structures. The evolution of embedding-based models continues to contribute to the flexibility and efficiency of topic modeling, offering novel insights into document representations in vector spaces. The transition from embedding-based models led to the rise of transformer-based models, with bidirectional encoder representations from transformers (BERT) in 2019 as a landmark.

BERT revolutionized natural language understanding by capturing context and relationships bi-directionally. Ongoing work involves fine-tuning BERT for specific tasks, such as topic modeling, leveraging its contextual understanding capabilities. This progression signifies a shift towards more powerful and context-aware models, as transformers, with their attention mechanisms, excel at capturing intricate relationships in data. The integration of transformer-based models, including BERT, into topic modeling frameworks represents a contemporary approach to enhancing the efficiency and performance of these models in capturing complex semantic structures. In this research, we explore several key methods that are used for topic modeling such as LDA, BERT, generative pre-trained transformers (GPT), and extra long-term memory networks (XLNet). A key discovery in this review is that when we apply BERT to the summaries created by XLNet, we consistently get very clear and meaningful topics. This discovery could change how we approach topic modeling and make it more useful in different areas.

This paper carefully examines and offers insights into how to make the most of these methods to uncover clear and meaningful topics in various applications. The main objectives of this study are given as follows. Firstly, it provides an overview of the field of topic modeling (TM), explaining what it is and why it is important. Then, it dives deep into different aspects of topic modeling like challenges, ways of doing it, where it is used, and how we measure its success. The study also gathers and presents important information, organizing and summarizing what we know about topic modeling. It also discussed the tools used in topic modeling. Lastly, it talks about the ways we measure how well topic modeling systems work. The remainder of this paper is structured so that section 2 shows the background of topic modeling, section 3 presents the literature review, section 4 presents the method, section 5 presents the results (comparison of models), and section 6 conclude the whole work.

2. BACKGROUND

2.1. Traditional modeling methods

2.1.1. Latent Dirichlet allocation

In the realm of text mining and data discovery, topic modeling stands as a powerful technique for unveiling latent relationships within data and text documents. Over the years, researchers have delved into this field, applying topic modeling to diverse domains, including software engineering, political science, medicine, and linguistics. Among the various methods available, LDA emerges as a cornerstone. This paper contributes to the existing body of knowledge by providing a comprehensive exploration of scholarly articles within the years 2003 to 2016, centered on LDA-based topic modeling. It not only sheds light on the evolution and current trends in this area but also delineates the intellectual landscape. Furthermore, this paper outlines the challenges encountered and introduces prominent tools and datasets essential to LDA-based topic modeling, making it a valuable resource for researchers and enthusiasts in this domain [9].

2.1.2. Non-negative matrix factorization

Non negative matrix factorization (NMF) [10] has emerged as a prominent technique for dimensionality reduction, demonstrating its prominence since its inception. With the incorporation of the nonnegativity constraint, NMF achieves a parts-based representation, which enhances the interpretability of the extracted features. This survey paper aims to provide a comprehensive overview of theoretical research on NMF over the past five years. It summarizes the principles, basic models, properties, and algorithms of NMF, as well as its modifications, extensions, and generalizations. The existing NMF algorithms are classified into four main groups: basic NMF (BNMF), constrained NMF (CNMF), structured NMF (SNMF), and generalized NMF (GNMF), with detailed analyses of their design principles, characteristics, relationships, and evolution. Additionally, the survey explores related work outside of NMF that offers insights or connections to NMF. Open research challenges are discussed, and several application areas of NMF are briefly outlined. This survey aims to provide an integrated and up-to-date framework for understanding NMF, with the goal of guiding future research endeavors in this field.

2.1.3. Latent semantic analysis

The paper referenced as [11] investigates LSA as a method for analyzing textual data, with a focus on its effectiveness in understanding relationships among terms within documents and across the corpus. LSA's application spans various domains, including intelligent information retrieval, search engines, and online news platforms, where it plays a crucial role in classification, clustering, summarization, and search tasks. Through the utilization of singular value decomposition (SVD), LSA endeavors to reveal the overall structure of documents and uncover latent connections within the dataset. Researchers employ LSA to gain deeper insights into term-document relationships, particularly in the context of analyzing research papers in natural language processing. The study demonstrates LSA's capacity to merge terms with similar meanings and identify subtle nuances, thereby streamlining document representation in a lower-dimensional space. The

paper referenced as [12] delves into the exploration of word embedding in the context of small corpora, particularly pertinent in the study of individual text production. A comparative analysis is conducted between skip-gram and LSA within this framework, focusing on their effectiveness in extracting semantic patterns from single-series dream reports. The findings reveal that LSA outperforms Skip-gram when applied to small training corpora in two distinct semantic tests. As a compelling case study, the paper demonstrates LSA's capacity to capture meaningful word associations within dream report series, even when dealing with a limited number of dreams or infrequent words. This highlights the potential of LSA as a valuable tool for unraveling word associations in the realm of dream reports, thereby contributing fresh perspectives to the timeless field of psychology research.

2.1.4. Probabilistic latent semantic analysis

Probabilistic latent semantic analysis (PLSA) treats topics as hidden variables and aims to uncover a probabilistic model that can effectively explain the observed data present in the document-term matrix. Essentially, PLSA seeks a model $(P(D, W))$ where, for any given document and word in the corpus, the probability $(P(D, W))$ corresponds with the relevant entry in the document-term matrix. This method merges the principles of LSA with probabilistic principles, offering a comprehensive framework for topic modeling in NLP. PLSA holds significant implications for uncovering the underlying structures within textual data, thus serving as a valuable tool in NLP research and applications. In paper [13], PLSA is a statistical technique widely employed in information retrieval and various other fields, known for its intuitive and robust results. Despite its effectiveness, PLSA is often criticized for the rigidity of its assumptions and the iterative nature of the expectation-maximization algorithm, which has led to divisions among researchers. In this manuscript, we provide an overview of PLSA, followed by a discussion on reformulations aimed at addressing its inherent problems. Special attention is given to approaches that connect PLSA with the singular value decomposition theorem, seeking to enhance its efficiency and effectiveness. Furthermore, we explore how PLSA serves as the foundation for other techniques, such as kernelization and probabilistic transfer learning. These extensions transform PLSA from a descriptive tool into an inferential one, thereby opening up new avenues for research and applications.

2.2. Deep learning approaches for topic modeling

In their study, Chai and Li [14] address the issue of interpretability in deep learning-based text mining, a challenge that has constrained the adoption and improvement of such models in the field of information systems (IS). They introduce a novel research framework called "neural topic embedding" that combines deep neural networks with topic modeling to create interpretable representations of textual documents. Through a preliminary experiment in fake review detection, their approach significantly enhances the state-of-the-art performance, as indicated by an 8 percent improvement in F1 score. This research marks a significant contribution to the IS community, offering a promising pathway for the future integration of state-of-the-art deep learning methods while ensuring interpretability in text mining applications.

2.2.1. Bidirectional encoder representations from transformers

Bidirectional encoder representations from transformers (BERTopic) [15] is an advanced topic modeling approach that harnesses the power of transformer-based deep learning models to identify topics in extensive text datasets. Its advantages include the generation of more coherent and interpretable topics compared to traditional methods like latent Dirichlet allocation. BERTopic excels in handling large text collections effectively and captures semantic relationships between words, resulting in highly accurate and meaningful topics. Moreover, it offers fine-grained control over the number of topics to be extracted and is amenable to fine-tuning specific text datasets for enhanced performance. In summary, BERTopic represents a significant advancement in topic modeling with a wide range of benefits for researchers and practitioners.

2.2.2. Generative pre-trained transformer

Generative pre-trained transformers, or GPT models [16], built on the transformer architecture, represent a pivotal advancement in artificial intelligence (AI). They empower applications like ChatGPT to generate human-like text and multimedia content, answer questions conversationally, and find applications across various industries. The significance of GPT lies in its transformative impact on AI research, introducing a new era of machine learning adoption. GPT models, due to their speed and scalability, can automate and enhance an array of tasks, from language translation and document summarization to content creation, website design, code generation, and even poetry composition. They are a driving force behind the pursuit of artificial general intelligence, offering the potential to elevate organizational productivity and transform applications and user experiences. This review paper explores the profound impact of GPT models in AI research and their wide-ranging applications.

2.2.3. Extra long-term memory networks

XLNet [17], which stands for “extra long-term memory networks,” is an advanced language model used in NLP. Unlike older models like GPT, XLNet introduces a new way of training called “permutation language modeling.” This helps it understand the order of words in a sentence better, capturing context in both directions. It combines two types of models, autoregressive (captures the context from left to right, as seen in traditional language models) and autoencoding (allows the model to leverage bidirectional context), allowing it to understand language more thoroughly. Similar to GPT, XLNet is versatile, excelling in tasks like text completion and translation. Its impact on AI research has been significant, overcoming some limitations of older models and contributing to advancements in natural language processing tasks. In simpler terms, XLNet is a powerful tool that understands language better, thanks to its innovative training approach, making it a key player in the development of artificial intelligence.

3. LITERATURE REVIEW

In this section, we review the work related to models based on embedding and the models in which topics are generated for specific parts of speech in a dataset. Embedding-based models have gained prominence for their ability to capture semantic relationships between words and phrases, providing a robust foundation for various natural language processing tasks. On the other hand, models that focus on generating topics for specific parts of speech, such as nouns or verbs, offer more fine-grained insights into the structure and thematic distribution of datasets. This dual focus on embedding techniques and targeted topic modeling highlights the evolving methodologies aimed at enhancing textual analysis and understanding.

3.1. Sentiment analysis and deep learning approaches in understanding textual data

In recent years a lot of research has been carried out that have shown that word embedding is quite useful for topic modeling. The review in this section is mainly consists of embedding based models. With the proliferation and expansion of online businesses have streamlined the process of obtaining direct customer feedback. This accessibility enables companies to identify their weaknesses and take measures to strengthen their market position. Additionally, sentiment analysis serves as a valuable tool for comprehending people’s emotions and attitudes toward their products [18]. While social media platforms have granted individuals a platform for free expression, they have also been associated with disturbing incidents. Hate speech on social media has emerged as a significant concern, posing a severe threat to people’s well-being. Sentiment analysis can play a crucial role in identifying and addressing hate speech on these platforms.

Various deep learning techniques, such as convolutional neural networks (CNN) and recurrent neural networks (RNN) can be employed for this purpose [19]. Stanford University conducted research utilizing emotion symbols to extract tweets, employing Twitter’s API, with each tweet representing either a positive or negative sentiment. To enhance the accuracy of classification models, several data preprocessing techniques were implemented. This research has the potential to automate the classification of tweets without the need for manual intervention. It should be noted that their sentiment analysis was conducted comprehensively without specific domain focus [20].

Twitter has become really important for figuring out what people think and feel because lots of different types of people use it. People use all sorts of methods to figure out what people are saying on Twitter, but sometimes it is not totally accurate because there are not always good sets of data to use, and understanding how language works can be tricky [21]. Sentiment analysis is useful in many areas, like entertainment and healthcare. People are studying it a lot, but sometimes it is hard to do research because there is not enough data. Even when there is data, getting permission and paperwork can take a long time, even months [22]. As social networking sites have become increasingly popular, people now tend to express their emotions online rather than confiding in their family members. This shift has sparked numerous research initiatives in the field of mental health, utilizing NLP techniques [23].

While data related to mental health remains highly confidential and often inaccessible due to agreements with data sources, recent technological advancements have introduced APIs like Tweepy and Twitter4j. These APIs enable the collection of online data for sentiment analysis, particularly from various social media platforms. Besides textual posts, many social media websites, such as Flickr and Instagram, facilitate the sharing and hosting of images. Analyzing these images, which reflect users’ moods and emotions, can provide valuable insights into their emotional states and sentiments [24]. Although social media is often associated with contributing to mental health issues like increased anxiety, it has also provided individuals with a platform to challenge stigmas and outdated beliefs, allowing them to express their thoughts without fear. Given the openness and freedom facilitated by social media, these platforms have a responsibility to address the growing mental health concerns. However, it is important to note that none of the existing research can guarantee the accurate prediction of all suicidal and non-suicidal posts on social media by different users [25]. The challenges arise due to issues inherent in natural language processing.

While achieving a 100% accurate result is difficult, sentiment analysis does provide valuable insights and, in some instances, can help prevent mishaps.

Beyond analyzing social media posts, there are other methods for assessing individual sentiment, such as classifying suicide notes. This process involves linguistic analysis, as individuals express their emotions uniquely in suicide notes. Differences in language, pronoun usage, and grammar make it challenging to accurately classify these notes [26]. Additional research efforts have also been dedicated to distinguishing genuine from fake suicide notes, employing a range of techniques from traditional approaches to cutting-edge machine learning methods.

One limiting factor in this research domain is the limited availability of sample data. The utilization of NLP and text mining in sentiment analysis is still evolving, which explains the challenges faced by major social media platforms in accurately identifying posts expressing negative sentiments [27]. In the early stages of sentiment analysis, the focus was primarily on individual words rather than comprehending the context in which these words appeared [28]. Subsequently, a WordNet-based approach was introduced for sentiment analysis, which involved calculating the distance between words in a text and the terms “good” or “bad” [29]. Some researchers even utilized Cosine distance to enhance accuracy. The adoption of bidirectional long short-term memory (Bi-LSTM) in sentiment analysis marked a shift towards considering context, making it a more effective method compared to others [30]. Dealing with extensive text and deciphering its meaning can be challenging, and one technique to better grasp the subjects within large collections of text is topic modeling [31]. Topic modeling has found applications in diverse fields such as political science, customer reviews, software engineering, medicine, and linguistics [32].

3.2. Topic modeling from traditional approaches to modern techniques

While there are numerous topic modeling algorithms available, optimizing and fine-tuning them is crucial for obtaining reliable results. A solid understanding of the underlying processes in these algorithms is necessary to determine which one best aligns with the specific objectives at hand [33]. Topic modeling continues to hold a pivotal role in the domains of machine learning and natural language processing, necessitating ongoing research efforts to enhance its utility across various domains. The landscape of topic modeling has undergone significant transformations since its inception. Among the earliest and most widely adopted models, LDA, originally introduced by Blei *et al.* in 2003 [3], remains a foundational and prevalent approach in the realm of probabilistic topic modeling. Reisenbichler and Reutterer [34] emphasized the application of topic models in marketing research, highlighting their efficacy in revealing hidden patterns in complex co-occurrence data. They noted the increasing interest among marketing scholars and practitioners in utilizing these versatile clustering techniques due to the availability of extensive datasets. However, they identified a significant gap in the literature. An absence of a comprehensive overview of this rapidly evolving field. To address this gap, the authors conducted a systematic review of 61 relevant papers and conceptual contributions, exploring various dimensions including data structures, model implementation, extensions, and performance evaluation. Their objective was to categorize and illuminate the contributions made in marketing research through topic models, showcasing advancements across marketing sub-domains. The authors review not only synthesized existing literature but also provided valuable insights into potential future research directions in marketing, with a focus on the utilization of topic modeling techniques.

Short texts exhibit less word co-occurrences than longer texts do, and word co-occurrences are crucial for topic modelling because existing models are based on word co-occurrences and fail when applied to short texts [35]. Meanwhile, Sivanandham *et al.* [36], in their work in 2021, employed a combination of techniques, including LDA topic modeling, NLP, time series analysis, and vector auto-regression, to analyze research trends and make predictions using topic modeling. It is worth noting that although their methods were well-detailed, their dataset only covered articles spanning approximately nine years, which may not provide entirely accurate results.

Grisales *et al.* [37] discussed the evolution and applications of topic modeling, a prominent natural language processing technique, by employing bibliometric analysis. It reveals that the USA and China are among the most productive countries in this field, with applications primarily focused on identifying sub-topics in short texts like social networks and blogs. The study underscores the versatility of topic modeling and its valuable role in systematically reviewing vast volumes of unstructured information, offering insights beneficial for researchers and academics engaged in various academic and research contexts. This study uses the LDA model to perform topic extraction on the abstracts of documents from 25 journals and 19 conferences marked. Yadav highlighted the significance of sentiment analysis in today's competitive world, showcasing its wide-ranging applications across various sectors, from healthcare to entertainment, corporate, and politics. The study focuses on analyzing headlines from India's leading news portal, the Times of India, employing both supervised and unsupervised techniques. In supervised sentiment analysis, the Bi-LSTM technique is employed, while unsupervised topic modeling is conducted using LDA and LSA, with LDA

outperforming LSA in terms of producing a more balanced distribution of topics. The research emphasizes the importance of choosing an appropriate number of topics based on dataset characteristics, providing valuable insights into topic selection methodologies beyond trial-and-error approaches [38]. This thesis project focuses on harnessing topic modeling techniques within the realm of natural language processing to extract valuable insights from customer calls to a European financial service provider. It aims to compare two prominent topic modeling algorithms, LDA and BERTopic, to categorize and analyze call content, ultimately enhancing the company's understanding of customer needs and facilitating more effective decision-making. LDA outperforms BERTopic in terms of topic quality and interpretability, indicating its superior ability to capture meaningful and coherent topics within the customer call data. The results demonstrate the potential for improving customer engagement, satisfaction, and tailored strategies for the company's benefit [39].

Word embedding is effective in identifying syntactically and semantically connected words, according to research, and they can be used to connect related words in a corpus [40]. Determining semantically relevant content in short texts can be challenging since they often have less word co-occurrences. Non-negative matrix factorization, which is based on pre-trained distributional vector representation, is used to solve this issue [41]. Patil *et al.* [42] discussed the use of topic modeling, specifically LDA as an unsupervised machine learning technique for discovering topics within a collection of documents. They emphasize that while LDA is typically applied to identify topics across multiple documents, it can also be valuable for single-document applications, where it effectively extracts keywords summarizing the core idea of the document concisely. This approach is particularly useful for text summarization, where the goal is to shorten a text document while retaining its essential points. The authors present an LDA model that identifies dominant topics within a text document and selects sentences reflecting these topics to create a human-readable summary.

According to Shi *et al.* [43], a neural word embedding is employed to determine the context of words in order to identify the themes of short texts. Similar to this, Wang *et al.* [44] employ word embedding's to identify additional information for integrating short text documents into sections of adjacent words. Word embedding are crucial for topic modelling to locate the target term's immediate context [45]. An alternate method of topic modelling with less run-time complexity is offered by pre-trained word embedding in combination with k-means clustering and TF-based re-ranking [46]. Hasan *et al.* [47] focused on topic modeling, a statistical data mining technique used to categorize diverse documents into coherent topics. The study evaluates the effectiveness of two prominent topic modeling methods when applied to Bangla news articles, with the goal of automating document categorization for recommendation systems and search applications. The methods under examination are LDA and LDA2vec, which combines LDA with Word2Vec.

Despite the inherent differences in syntax between Bangla and English, LDA demonstrates its efficacy in extracting meaningful topics from Bangla news documents. To facilitate testing and practical implementation, a novel technique is devised for identifying topics in non-factorized documents by LDA and LDA2vec. Remarkably, the results reveal that LDA2vec outperforms LDA, achieving an impressive accuracy rate of 85.66% in identifying topics within test documents, compared to LDA's accuracy of 62.45%. As such, LDA2vec emerges as the more reliable choice for real-world applications. Qi and He [48], presented an off-topic detection algorithm that enhances accuracy and efficiency in English composition-assisted review systems. It combines LDA and Word2Vec to model and train documents, leveraging semantic relationships between document topics and words to calculate probability-weighted sums for each topic and its feature words. The algorithm identifies off-topic compositions through a set threshold, outperforming the vector space model in experimental results, achieving higher accuracy and an F value exceeding 88%. This algorithm holds promise for intelligent off-topic detection, especially in English composition teaching contexts.

Grootendorst in study [49] discussed the BERTopic, a topic modeling approach that goes beyond traditional methods by leveraging state-of-the-art language models and introducing a class-based TF-IDF procedure for creating coherent topic representations. BERTopic operates in several steps: it first generates document embedding using pre-trained transformer-based language models, then clusters these embedding, and finally produces topic representations using the class-based TF-IDF approach. This separation of tasks provides flexibility and ease of use in the model. The paper presents a comprehensive analysis of BERTopic, including evaluations with classical topic coherence measures and considerations of running times. The results of experiments indicate that BERTopic effectively captures coherent language patterns and consistently performs well across various tasks, making it a competitive and stable option in the field of topic modeling. Cygan [50] highlighted the significance of browsing history, which often contains a wealth of intellectual exploration recorded as URLs, titles, and timestamps. However, the tabular format of this data makes it challenging to explore and extract meaningful insights. The author employs Sentence-BERT (SBERT) to create expressive embedding's that facilitate topic modeling within browsing history data.

Qualitative analysis reveals that topic clusters generated from Sentence-BERT (SBERT) web page embeddings outperform those created using Doc2Vec-based document embedding. Despite a lower topic coherence score, SBERT-based embeddings excel in organizing a diverse range of documents without the need for prior training, unlike Doc2Vec, which requires training on the input dataset. SBERT embedding's capture semantic concepts effectively, while Doc2Vec tends to prioritize structural similarity over semantic meaning. Overall, this suggests that training on the input dataset for document topic modeling may not be ideal, as it can lead to an overemphasis on structural composition rather than semantic content. In Groot *et al.* [51], discussed the application of topic modeling, particularly latent Dirichlet allocation, and the challenges it faces when dealing with short texts from various domains. They explore the performance of the BERTopic algorithm, a state-of-the-art approach, on short multi-domain texts and find that it outperforms latent Dirichlet allocation in terms of topic coherence and diversity. Additionally, they analyze the performance of the HDBSCAN clustering algorithm used by BERTopic and identify an issue where it categorizes a significant portion of the documents as outliers, which hinders further analysis. To address this problem, they replace HDBSCAN with k-means clustering.

Liu *et al.* [52] conducted a study focusing on the utilization of summary generation and topic modeling to discern factors influencing vaccine attitudes across regions. They gathered tweets concerning Sinovac, AstraZeneca, and Pfizer vaccines, applying BERTopic clustering to categorize tweets into topics. Additionally, they utilized contrastive learning (CL) to produce summaries for each topic. The study found that BERTopic outperforms latent Dirichlet allocation and identified three consistent factors affecting vaccine attitudes: vaccine-related factors, health system-related factors, and individual social attributes. The research concludes that deep learning methods effectively reveal these factors, assisting policymakers and medical institutions in formulating more efficient vaccination strategies. Zankadi *et al.* [53] discussed the crucial role of learners' interests in the realm of education, specifically focusing on massive open online courses (MOOCs). It highlights the significant potential of nurturing and aligning learners' interests to enhance their overall MOOC learning experiences. The study delves into the concept that learners often reveal their genuine interests and preferences on social media platforms through user-generated content, which contains valuable hidden insights. The primary research objective is to identify and extract these topical interests from learners' social media posts, thereby enriching their course preferences within the MOOC context. To achieve this, the study employs NLP techniques and leverages various topic modeling methods, including latent Dirichlet allocation, latent semantic analysis, and BERTopic. Remarkably, the experimental findings underscore that BERTopic outperformed the other models when applied to the dataset collected for this study.

4. METHOD

This study focuses on evaluating the performance of transformer-based language models for text summarization using benchmark datasets. A structured approach is adopted, leveraging the Twenty Newsgroups dataset and an articles dataset to ensure diversity and representativeness in the evaluation. The datasets were carefully prepared and analyzed to facilitate meaningful comparisons and insights.

4.1. Selection of dataset

Two distinct datasets were utilized in this study to evaluate the performance of transformer-based language models for text summarization. The first is the Twenty Newsgroups dataset, a widely used collection of 18,846 documents categorized into 20 different newsgroups. This dataset, fetched using the `fetch_20newsgroups` function from the `scikit-learn` library, provides a rich and diverse textual base for natural language processing tasks. The second dataset, referred to as the Articles dataset, comprises textual data extracted from the file `articles.xlsx`, sourced from KDnuggets. This dataset is notable for its structured and comprehensive content, offering a complementary perspective to the unstructured nature of the Twenty Newsgroups dataset. Together, these datasets provide a balanced framework for evaluating text summarization models across varied textual contexts.

4.1.1. Twenty newsgroups dataset

The Twenty Newsgroups dataset, is a commonly used text dataset for various natural language processing tasks. The 20 Newsgroups dataset contains a total of 18,846 newsgroup documents, which are often referred to as "articles" organized as a collection of 20 different newsgroups categories. This dataset is fetched from the `sci-kit-learn` library using the `fetch_20newsgroups` function and consists of a total of 20 newsgroup categories, each with multiple articles.

4.1.2. Articles dataset

In this section, we describe the process of loading ‘*articles.xlsx*,’ which was obtained from KDnuggets and serves as the primary source of information for our research. This dataset is characterized by its size and structure, which are key factors in understanding the scope of the data. The dataset exhibits the following characteristics.

In Table 1, each row corresponds to a unique research article, and the columns provide information about the numerical identifier, title, journal, publication year, and abstract of each article. This structured representation facilitates easy retrieval and analysis of the articles for summarization tasks. By organizing the data in a tabular format, the dataset ensures clarity and consistency, enabling efficient preprocessing and feature extraction.

Table 1. Characteristics of articles.xlsx

Unnamed: 0	Title	Journal	Year	Abstract
1	[Title 1]	[Journal 1]	20xx	[Abstract 1]
2	[Title 2]	[Journal 2]	20xx	[Abstract 2]
3	[Title 3]	[Journal 3]	20xx	[Abstract 3]
....
6000	[Title 6000]	[Journal 6000]	20xx	[Abstract 6000]

4.2. Text preprocessing

After the selection of the dataset, the following text preprocessing steps are performed to remove noise, stop words, and redundancies from the dataset. These preprocessing steps ensure that the text is clean and standardized, improving the quality of input for the summarization models. By eliminating irrelevant elements, the process enhances the focus on meaningful content, contributing to more accurate and concise summaries.

4.2.1. Data preprocessing for LDA

In the preprocessing workflow for LDA, we follow a series of steps to prepare the text data for topic modeling. First, we leverage the spaCy library with the ‘*en_core_web_sm*’ model, disabling the parser and named entity recognizer for efficiency. Additionally, we utilize natural language toolkit (NLTK) stop words list. The primary preprocessing steps are as follows:

- Text cleaning: We begin by removing stop words and short tokens, while also eliminating letter accents from the text data.
- Bi-gram and tri-gram implementation: Next, we build bi-grams and tri-grams to capture meaningful word combinations in the text.
- Lemmatization: We then apply lemmatization to the tokens and filter for specific part-of-speech tags (NOUN, ADJ, VERB, ADV).
- Final cleanup: After lemmatization, we remove stop words and short tokens once more, ensuring that the processed text data is clean and ready for topic modeling.

4.2.2. Preprocessing for BERTopic

The following preprocessing steps are executed:

- Text tokenization: The text data is tokenized, converting it into individual words units to make input data suitable for BERTopic. Libraries like NLTK and spaCy are commonly employed for this purpose.
- Stop words removal: Common stop words, such as ‘and,’ ‘the,’ and ‘in,’ are eliminated from the text which reduces noise and focuses on more meaningful words.
- UMAP dimension reduction: uniform manifold approximation and projection (UMAP) is applied to the text data. UMAP is a dimensionality reduction technique that helps in reducing the complexity of the dataset while preserving its intrinsic structure. It is configured with specific parameters such as the number of neighbors, components, minimum distance, and metric. This dimension reduction aids in optimizing the topic modeling process.

4.2.3. Preprocessing for generated summaries from GPT3

The preprocessing of summaries generated from GPT plays a crucial role in refining and organizing the text data for subsequent analysis. The following preprocessing steps are performed:

- Tokenization: initially, the text is tokenized, breaking it down into individual words. Each word is converted to lowercase to ensure consistency.

- b. Punctuation and non-alphanumeric removal: punctuation and non-alphanumeric characters are removed from the text. This step aids in eliminating noise and simplifying the content.
- c. Stop words removal: common stop words in the English language are removed from the text which enhances the focus on essential content.
- d. Lemmatization: lemmatization is applied to the words in the text to ensure uniformity and improve the quality of the text.
- e. Reconstruction: after tokenization, removal of punctuation and stop words, and lemmatization, the words are reconstructed into a sentence format. Commas are used to separate the lemmatized words, providing a clean, structured format for each summary.

4.2.4. Preprocessing steps of XLNet

The text preprocessing workflow involves a series of steps performed before and after generating summaries to prepare the data for analysis and ensure the quality of the generated outputs.

- a. Before generating summaries
 - Text tokenization: After loading the data the text data is tokenized into words using the XLNet model’s tokenizer.
 - Text processing with the model: The code generates summaries for each document using the XLNet model. It tokenizes the input text, adjusts the length, and generates the summary.
- b. After generating summaries
 - Initializing preprocessed summaries list: The “preprocessed summaries” is initialized to store the preprocessed versions of the generated summaries.
 - Stop words Removal: The NLTK stop words are removed from the generated summaries to eliminate common words that do not add significant meaning.
 - Re-tokenization: The preprocessed summaries are tokenized again into words using NLTK’s *word_tokenize* function.
 - Punctuation removal: Punctuation and non-alphanumeric characters are removed from the preprocessed summaries.
 - Word lemmatization: Words in the preprocessed summaries are lemmatized to reduce them to their base form.
 - Joining Words: The preprocessed words are joined back into a sentence with commas to create a single string for each preprocessed summary.

The preprocessing steps before generating summaries are related to generating the summaries themselves, while the preprocessing steps after generating summaries focus on cleaning and refining the generated summaries for further analysis or presentation

4.3. Evaluation metrics

We study the two models quantitatively. We measure the quality of the topics based on two metrics used in the study [54] topic coherence and topic diversity. Both these metrics are also used in the current study. The topic coherence is the measure that shows how frequently words co-occur in a dataset [55]. The higher the topic coherence the more interpretable is the topic model. According to [55], the coherence of a topic can be calculated as:

$$TC = \sum_{M=2}^m \sum_{l=1}^{m-1} \log \frac{|D(w_m^k, w_l^k)| + 1}{D(w_l^k)}$$

Whereas $w^k, w_1^k, w_2^k, w_3^k, \dots, w_m^k$ is the top M words under the topic k sorted by probability in descending order. $|D(w_m^k, w_l^k)|$ is the number of documents in a dataset containing both words w_m^k and w_l^k . In other words, the high the coherence score the higher will be the mutual information between words. All the models which show high coherence score generate highly interpretable topics. While topic diversity is the measure that shows that how unique words are inside the topics, the value of redundant topics will be near to 0 whereas the value near to 1 indicate unique topics. The main idea behind the diversity is the percentage of unique words in all the topics, the value close to 0 indicates more redundant topics whereas the value close to 1 indicate varied topics. The basic idea behind topic coherence is that the words happen in same document will have high coherence score. The topic quality is the result of multiplication of topic coherence and topic diversity.

5. RESULTS

Tables 2 to 5 show that the coherence and diversity of various models across the datasets (44 summaries from 20 newsgroups and articles). These tables provide a detailed comparison of the models' abilities to maintain logical flow and introduce varied perspectives in the generated summaries. The results highlight the strengths and weaknesses of each model, offering insights into their suitability for different types of textual data.

As shown in Tables 2 and 3, it is evident that BERT and XLNet-BERT lead in terms of topic coherence (TC), showcasing the highest scores among the listed models. Conversely, for topic diversity (TD), GPT-LDA stands out with the highest score, closely followed by XLNet-BERT. Considering both coherence and diversity, XLNet-BERT emerges as the top performer overall, boasting one of the highest TC scores and a commendable TD score. However, it is worth noting that BERT also excels, particularly in TC. If prioritizing diversity, GPT-LDA presents a compelling option. Therefore, the ultimate choice of the best performer hinges on the specific emphasis placed on coherence and diversity. Overall, for a balanced consideration of both aspects, XLNet-BERT appears to be the most suitable choice.

Table 2. Topic coherence (TC) of 20 newsgroups

Models comparison	
Models	Coherence Score
LDA	0.582948
BERT	0.69312
GPT-LDA	0.52703
XLNet-LDA	0.473952173
GPT-BERT	0.301965419
XLNet-BERT	0.696075298

Table 3. Topic diversity (TD) of 20 newsgroup

Models	Diversity score
LDA	1
BERT	0.69312
GPT-LDA	0.95442
XLNet-LDA	0.372309172
GPT-BERT	0.5589
XLNet-BERT	0.6656

Table 4. Topic coherence (TC) of articles dataset

Models Comparison	
Models	Coherence Score
LDA	0.408743
BERT	0.50338
GPT-LDA	0.279367
XLNet-LDA	0.32899
GPT-BERT	0.49588
XLNet-BERT	0.50438

Table 5. Topic diversity (TD) of articles dataset

Models	Diversity Score
LDA	1
BERT	0.3873
GPT-LDA	0.392484
XLNet-LDA	0.22994
GPT-BERT	0.5045
XLNet-BERT	0.6808

As shown in the Table 3, when evaluating the topic coherence (TC) scores for the models assessed on the articles dataset, BERT emerges as the frontrunner with a notable TC score of 0.50338, closely trailed by XLNet-BERT, which achieves a comparable score of 0.50438. On the other hand, while transitioning to topic diversity (TD), XLNet-BERT demonstrates its prowess by securing the highest TD score of 0.6808 among all models. This signifies XLNet-BERT's ability to not only maintain coherence but also to encompass a broader range of diverse topics within the dataset. Considering both coherence and diversity metrics, XLNet-BERT emerges as the optimal performer, balancing high TC scores akin to BERT while exhibiting superior diversity.

Therefore, XLNet-BERT stands out as the preferred choice for effectively capturing coherent and diverse topics within the articles dataset. Tables 6 to 9 show that the coherence and diversity of various models across all datasets (44 summaries from 20 newsgroups and 116 from articles dataset).

Table 6. Topic coherence (TC) of 20 newsgroup

Models	Coherence score
LDA	0.582946
BERT	0.69312
GPT-LDA	0.527030388
XLNet-LDA	0.473952173
GPT-BERT	0.56016
XLNet-BERT	0.69607

Table 7. Topic diversity (TD) of 20 newsgroups

Models	Diversity score
LDA	1
BERT	0.8978
GPT-LDA	0.954418454
XLNet-LDA	0.372309172
GPT-BERT	0.5589
XLNet-BERT	0.6556

Table 8. Topic coherence (TC) of articles

Models	Coherence score
LDA	0.48743
BERT	0.53388
GPT-LDA	0.3342
XLNet-LDA	0.29829244
GPT-BERT	0.56026
XLNet-BERT	0.69011

Table 9. Topic diversity (TD) of articles

Models	Diversity score
LDA	1
BERT	0.878
GPT-LDA	0.83557
XLNet-LDA	0.37230917
GPT-BERT	0.6906
XLNet-BERT	0.6956

Among the models evaluated as shown in Table 6 and 7, BERT and XLNet-BERT demonstrate exceptional performance in capturing coherent topics, with BERT achieving the highest topic coherence (TC) score of 0.69312 and XLNet-BERT closely following with a score of 0.69607 for the 20 newsgroups dataset. However, when considering topic diversity (TD) in the articles dataset, GPT-LDA attains the highest TD score of 0.954418454, closely trailed by XLNet-BERT with a score of 0.6556. Taking into account both coherence and diversity metrics, XLNet-BERT emerges as the best performer overall. Not only does it achieve a high TC score comparable to BERT for the 20 newsgroups dataset, but it also demonstrates a relatively high TD score among the models listed for the articles dataset. Thus, XLNet-BERT stands out as the preferred choice for effectively capturing coherent topics in the 20 newsgroups dataset while also encompassing a diverse range of topics within the articles dataset.

Based on the provided data, XLNet-BERT emerges as the top performer for the articles dataset. XLNet-BERT achieves the highest topic coherence (TC) score of 0.69011 as given in Table 8, indicating its effectiveness in capturing coherent topics within the dataset. Additionally, XLNet-BERT demonstrates a relatively high topic diversity (TD) score of 0.6956, as shown in Table 9, showcasing its ability to encompass a diverse range of topics. While GPT-LDA achieves the highest TD score of 0.83557 as given in table but have less coherence score, XLNet-BERT's strong performance in both TC and TD metrics makes it the preferred choice overall. Therefore, XLNet-BERT stands out as an effective model for effectively capturing coherent and diverse topics within the articles dataset. Table 10 shows the overall best models in terms of coherence and diversity.

Table 10. Overall models performance

Dataset	Best model for coherence	Best model for diversity
20 Newsgroups	XLNet-BERT	GPT-LDA
Articles	XLNet-BERT	GPT-LDA, LDA, XLNet-BERT

Hybrid models, as explored in this comparative analysis, represent a fusion of distinct natural language processing techniques to leverage the strengths of individual models for enhanced performance. GPT-LDA and XLNet-LDA combine the capabilities of generative pre-trained transformers (GPT and XLNet) with LDA, a traditional topic modeling algorithm. In the context of this study, GPT is employed for generating summaries, which are subsequently analyzed by LDA to extract topics. Similarly, XLNet, a transformer model, is utilized for summarization in conjunction with LDA for topic extraction. The GPT-BERT and XLNet-BERT hybrids integrate generative pre-trained transformers with BERT, a bidirectional transformer model renowned for contextualized embeddings. These hybrid models, in essence, harness the proficiency of transformer-based models in understanding context and generating coherent summaries, coupled with the topic extraction capabilities of LDA or BERT. The motivation behind employing hybrid models lies in the belief that combining the strengths of different algorithms can potentially yield more robust and nuanced results. However, the observed variations in topic coherence and topic diversity across the hybrid models in the presented tables indicate that the synergy between summarization and topic modeling is a complex interplay, with each hybrid model exhibiting unique strengths and weaknesses.

This study sheds light on the intricacies of such hybrid approaches, providing valuable insights into their performance in the context of topic modeling and summarization tasks. XLNet, a powerful transformer-based language model, stands out for its ability to comprehend contextual relationships within text. Developed with a bidirectional context mechanism, XLNet captures dependencies between words both from the left and right, allowing for a nuanced understanding of language. This bidirectional approach enhances its performance across a spectrum of natural language processing tasks. When applied to text summarization, XLNet proves particularly adept at generating coherent and meaningful summaries. Leveraging its contextual understanding, XLNet distills essential information from input text, providing concise and comprehensive summaries. The model's capacity to consider the entirety of a document and capture long-range dependencies makes it well-suited for summarization tasks.

To further augment the utility of these summaries, the XLNet-BERT approach is employed. In this two-step process, XLNet first generates summaries, and then BERT comes into play to extract topics from these summaries. BERT, a formidable language representation model, excels in discerning key topics by comprehending bidirectional relationships and contextual nuances between words. This fusion of XLNet and BERT, referred to as XLNet-BERT, delivers competitive results in both coherence and diversity across diverse datasets. This indicates that the model not only produces summaries with logical flow but also extracts a diverse array of meaningful topics from those summaries.

Consequently, XLNet-BERT emerges as the preferred model for tasks requiring the extraction of nuanced and varied insights from textual data and thus shows as the best overall performer in generating meaningful and diverse topics across the provided datasets. XLNet-BERT emerges as the top performer for coherence in both the 20 Newsgroups and Articles datasets, exhibiting the highest scores among the models considered. On the other hand, GPT-LDA demonstrates superior diversity performance, particularly excelling in the 20 Newsgroups dataset. It achieves the highest diversity score in this dataset, showcasing its ability to capture a wide range of topics effectively. Furthermore, in the Articles dataset, GPT-LDA also leads in diversity, closely followed by LDA and XLNet-BERT. Overall, XLNet-BERT consistently showcases strong coherence across all datasets, while GPT-LDA stands out for its diversity capabilities, especially in the 20 newsgroups dataset. Additionally, LDA demonstrates noteworthy performance in diversity within the articles dataset.

6. CONCLUSION

In this paper, we explored the world of topic modeling. Word embedding is being widely used due to its efficiency as they are able to identify the semantic relation between words and represent them as vectors of real numbers. Word embedding is faster than other baseline techniques, also word embedding have lower dimensionality so they require less memory and computational resources. This study thoroughly explored various topic modeling models, ranging from traditional methods like LDA to advanced transformer-based approaches such as BERT, as well as hybrid models combining GPT, XLNet, and BERT. We focused on important metrics like topic coherence (TC) and topic diversity (TD) to better understand the semantic relationships between words and improve the precision of topic modeling. To compare these models, we applied them to two different datasets. The results analysis revealed specific strengths and

weaknesses across models. BERT demonstrated high efficiency in capturing semantic relations, leading to high topic coherence. On the other hand, GPT-LDA excelled in generating diverse topics, showcasing its ability to achieve high topic diversity. When summarizing the overall model performance, XLNet-BERT consistently outperformed other models in terms of topic coherence. This highlights the success of the hybrid approach, combining XLNet for summary generation and BERT for topic extraction, resulting in more coherent and meaningful topics. Notably, XLNet-BERT maintained robust performance across both datasets, emphasizing its capability to generate diverse topics. Our proposed research establishes the superiority of the hybrid model XLNet-BERT, providing a well-balanced synthesis of coherence and diversity. These findings offer valuable insights into the field of topic modeling, emphasizing the importance of simultaneously considering both aspects for a nuanced understanding of textual data. While promising, further investigation is imperative to refine models and explore additional dimensions, promising avenues for future research to achieve heightened accuracy and understanding in the evolving landscape of topic modeling. Additionally, considering emerging techniques and incorporating advancements in transformer-based models may contribute to more nuanced and sophisticated topic modeling methodologies.

ACKNOWLEDGMENT

The authors extend their appreciation to the Arab Open University for funding this work through AOU research fund No. (AOUKSA524008).




REFERENCES

- [1] C. U. Ghanshyambhai, "Optimizing topic coherence in the Gujarati text topic modeling: a relevant words-based approach," Ph.D. dissertation, Department of Computer Science and Engineering, Faculty of Technology, Gujarat Technological University.
- [2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Journal of Machine Learning Research*, vol. 3, no. 4–5, The MIT Press, 2003, pp. 993–1022, doi: 10.7551/mitpress/1120.003.0082.
- [4] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 1999, pp. 50–57, doi: 10.1145/312624.312649.
- [5] J. Zhu, A. Ahmed, and E. P. Xing, "MedLDA: Maximum margin supervised topic models for regression and classification," in *ACM International Conference Proceeding Series*, Jun. 2009, vol. 382, pp. 1257–1264, doi: 10.1145/1553374.1553535.
- [6] J. Zhu, N. Chen, H. Perkins, and B. Zhang, "Gibbs max-margin topic models with data augmentation," *Journal of Machine Learning Research*, vol. 15, pp. 1073–1110, 2014.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Neural information processing systems*, vol. 1, pp. 1–9, 2006.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, Jan. 2013.
- [9] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, Nov. 2019, doi: 10.1007/s11042-018-6894-4.
- [10] W. S. Chen, Q. Zeng, and B. Pan, "A survey of deep nonnegative matrix factorization," *Neurocomputing*, vol. 491, pp. 305–320, Jun. 2022, doi: 10.1016/j.neucom.2021.08.152.
- [11] P. Kherwa and P. Bansal, "Latent semantic analysis: An approach to understand semantic of text," in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017*, Sep. 2018, pp. 870–874, doi: 10.1109/CTCEEC.2017.8455018.
- [12] E. Altszyler, M. Sigman, S. Ribeiro, and D. F. Slezak, "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database," *arXiv:1610.01520*, Oct. 2016, doi: 10.1016/j.concog.2017.09.004.
- [13] F. Pau, G. B. Pablo, and G. Scholar, "Revisiting the probabilistic latent semantic analysis: The method, its extensions and its algorithms," *Prep*, Sep. 2023, doi: 10.20944/preprints202309.0293.v1.
- [14] Y. Chai and W. Li, "Towards deep learning interpretability: A topic modeling approach," *ICIS 2019 Proceedings*, 2019.
- [15] D. Napolitano, "Log analysis: Topic modeling applications on fine-features data processing system," Ph.D. dissertation, Politecnico di Torino, 2022.
- [16] G. Yenduri *et al.*, "GPT (generative pre-trained transformer) - A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024, doi: 10.1109/ACCESS.2024.3389497.
- [17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [18] S. Brody and M. Lapata, "Bayesian word sense induction," in *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 2009, pp. 103–111, doi: 10.3115/1609067.1609078.
- [19] M. E. Sunil and S. Vinay, "Kannada sentiment analysis using vectorization and machine learning," in *Sentimental Analysis and Deep Learning*, Springer Singapore, 2022, pp. 677–689, doi: 10.1007/978-981-16-5157-1_53.
- [20] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford University*, vol. 1, no. 12, pp. 1–6, 2009.
- [21] T. Sahni, C. Chandak, N. R. Chedeti, and M. Singh, "Efficient Twitter sentiment classification using subjective distant supervision," in *2017 9th International Conference on Communication Systems and Networks, COMSNETS 2017*, Jan. 2017, pp. 548–553, doi: 10.1109/COMSNETS.2017.7945451.
- [22] B. Gaye, D. Zhang, and A. Wulamu, "A tweet sentiment classification approach using a hybrid stacked ensemble technique," *Information*, vol. 12, no. 9, p. 374, Sep. 2021, doi: 10.3390/info12090374.




- [23] R. Behera, R. N. Behera, M. Roy, and S. Dash, "A novel machine learning approach for classification of emotion and polarity in sentiment140 dataset," in *3rd International Conference on Business & Information Management (ICBIM-2016)*, NIT, Durgapur, India, 2015.
- [24] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks," *Procedia Computer Science*, vol. 113, pp. 65–72, 2017, doi: 10.1016/j.procs.2017.08.290.
- [25] Z. Xu, V. Pérez-Rosas, and R. Mihalcea, "Inferring social media users' mental health status from multimodal information," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, pp. 6292–6299.
- [26] S. T. Rabani, Q. R. Khan, and A. M. Ud Din Khanday, "Detection of suicidal ideation on Twitter using machine learning & ensemble approaches," *Baghdad Science Journal*, vol. 17, no. 4, pp. 1328–1339, Dec. 2020, doi: 10.21123/bsj.2020.17.4.1328.
- [27] A. M. Schoene, G. Lacey, A. P. Turner, and N. Dethlefs, "Dilated LSTM with attention for classification of suicide notes," in *LOUHI@EMNLP 2019 - 10th International Workshop on Health Text Mining and Information Analysis, Proceedings*, 2019, pp. 136–145, doi: 10.18653/v1/d19-6217.
- [28] A. C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran, "Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing," *Scientific Reports*, vol. 8, no. 1, May 2018, doi: 10.1038/s41598-018-25773-2.
- [29] M. Taboada, J. Brooke, M. Tofloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011, doi: 10.1162/COLI_a_00049.
- [30] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using WordNet to measure semantic orientations of adjectives," in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, 2004, pp. 1115–1118.
- [31] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.
- [32] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: a survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–35, Sep. 2022, doi: 10.1145/3462478.
- [33] H. Jelodar *et al.*, "Latent dirichlet allocation (LDA) and topic modeling," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.
- [34] M. Reisenbichler and T. Reutterer, "Topic modeling in marketing: recent advances and research opportunities," *Journal of Business Economics*, vol. 89, no. 3, pp. 327–356, Apr. 2019, doi: 10.1007/s11573-018-0915-7.
- [35] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1–16, Jul. 2020, doi: 10.4108/eai.13-7-2018.159623.
- [36] S. Sivanandham, A. Sathish Kumar, R. Pradeep, and R. Sridhar, "Analysing research trends using topic modelling and trend prediction," *Soft Computing and Signal Processing*, pp. 157–166, 2021. Accessed: May 12, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-33-6912-2_15
- [37] A. M. Grisales, S. Robledo, and M. Zuluaga, "Topic modeling: Perspectives from a literature review," *IEEE Access*, vol. 11, pp. 4066–4078, 2023, doi: 10.1109/ACCESS.2022.3232939.
- [38] V. Yadav and S. Shakya, "Sentiment analysis and topic modeling on news headlines," *Journal of Ubiquitous Computing and Communication Technologies*, vol. 4, no. 3, pp. 204–218, 2022, doi: 10.36548/jucct.2022.3.008.
- [39] H. Axelborn and J. Berggren, "Topic modeling for customer insights: A comparative analysis of LDA and BERTopic in categorizing customer calls," M.S. thesis, Umea University, 2023.
- [40] X. Li, A. Zhang, C. Li, L. Guo, W. Wang, and J. Ouyang, "Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings," *Computer Journal*, vol. 62, no. 3, pp. 359–372, May 2019, doi: 10.1093/comjnl/bxy037.
- [41] F. Yi, B. Jiang, and J. Wu, "Topic modeling for short texts via word embedding and document correlation," *IEEE Access*, vol. 8, pp. 30692–30705, 2020, doi: 10.1109/ACCESS.2020.2973207.
- [42] C. Patil, P. Wayangankar, P. Yadav, and S. Sharm, "Review of topic modeling and summarization," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 3, 2022, doi: 10.32474/IRJET.2022.09.03.0146.
- [43] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 2018, pp. 1105–1114, doi: 10.1145/3178876.3186009.
- [44] J. Wang, L. Chen, L. Qin, and X. Wu, "ASTM: an attentional segmentation based topic model for short texts," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Nov. 2018, pp. 577–586, doi: 10.1109/ICDM.2018.00073.
- [45] G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang, "A correlated topic model using word embeddings," in *IJCAI International Joint Conference on Artificial Intelligence*, Aug. 2017, pp. 4207–4213, doi: 10.24963/ijcai.2017/588.
- [46] L. Shi, G. Cheng, S. Xie, and G. Xie, "A word embedding topic model for topic detection and summary in social networks," *Measurement and Control*, vol. 52, no. 9–10, pp. 1289–1298, Nov. 2019, doi: 10.1177/0020294019865750.
- [47] M. Hasan, M. M. Hossain, A. Ahmed, and M. S. Rahman, "Topic modelling: A comparison of the performance of latent Dirichlet allocation and LDA2vec model on Bangla newspaper," in *2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019*, Sep. 2019, pp. 1–5, doi: 10.1109/ICBSLP47725.2019.202047.
- [48] Y. Qi and J. He, "Application of LDA and word2vec to detect English off-topic composition," *PLOS ONE*, vol. 17, no. 2, Feb. 2022, doi: 10.1371/journal.pone.0264552.
- [49] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv:2203.05794*, Mar. 2022.
- [50] N. Cygan, "Sentence-BERT for interpretable topic modeling in web browsing data." Technical Report CS224N, Department of Computer Science, Stanford University, 2021.
- [51] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on generalizability of BERTopic on multi-domain short text," *arXiv:2212.08459*, Dec. 2022.
- [52] Y. Liu, J. Shi, C. Zhao, and C. Zhang, "Generalizing factors of COVID-19 vaccine attitudes in different regions: A summary generation and topic modeling approach," *Digital Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231188852.
- [53] H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," *Education and Information Technologies*, vol. 28, no. 5, pp. 5567–5584, Nov. 2023, doi: 10.1007/s10639-022-11373-1.
- [54] S. Li, R. Pan, H. Luo, X. Liu, and G. Zhao, "Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling," *Knowledge-Based Systems*, vol. 218, Apr. 2021, doi: 10.1016/j.knsys.2021.106827.
- [55] J. Jiang, "Modeling syntactic structures of topics with a nested HMM-LDA," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Dec. 2009, pp. 824–829, doi: 10.1109/ICDM.2009.144.

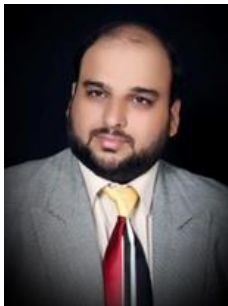
BIOGRAPHIES OF AUTHORS






Ayesha Riaz    received the BS degree in computer science from the Abdul Wali Khan University, Mardan, Pakistan, in 2021. Currently, she is pursuing her MS degree in software engineering from University of Engineering and Technology, Mardan, Pakistan. Her research interests include artificial intelligence, blockchain technology, software engineering, and cyber security. She can be contacted at email: aashiikhan.101@gmail.com.






Omar Abdulkader    received his B.E. degree in computer science from King Abdulaziz University, Jeddah, KSA, in 2002. He received his M.Sc. degree “Traffic parameter extracting using image process” from King Abdulaziz in 2010. He received his Ph.D. degree “Analytical cybersecurity model based on lightweight cryptographic for IoT” from King Abdulaziz University in 2019, Jeddah, KSA. His current research interests include cybersecurity, IoT, M2M, artificial intelligence, blockchain, cloud computing, location-based services, and machine learning. He can be contacted at email: o.abdulkader@arabou.edu.sa.



Muhammad Jawad Ikram    received the B.E. degree in computer systems engineering from the University of Engineering and Technology, Peshawar, Pakistan, in December 2010, the M.Sc. degree with distinction in networks and performance engineering from the University of Bradford, U.K., in December 2012, and the Ph.D. degree from King Abdulaziz University (KAU), Jeddah, Saudi Arabia, in March 2018. From August 2018 till August 2023, he served as an assistant professor of computer science at various universities including International Islamic University (IIU), Islamabad, Pakistan (2018-2019), Riphah International University, Pakistan (2019), PMAS Arid Agriculture University, Rawalpindi, Pakistan (2019-2020), Jeddah International College, Saudi Arabia (2020-2023). Since September 3, 2023, he has been working as an assistant professor at The Arab Open University, Saudi Arabia. He has published his research in a number of refereed journals and conferences. His recent research interests include machine learning, energy-aware algorithms, exascale computing, HPC, GPU computing, game theory, internet of things, and performance modeling. He can be contacted on his email at m.ikram@arabou.edu.sa.



Sadaqat Jan    currently serves as a professor in the Department of Software Engineering at the University of Engineering and Technology Mardan, Pakistan. He received his Ph.D. degree from Brunel University, UK, in 2011. He has authored a number of papers in preferred Journals and conferences. His research mainly focuses on semantic web, text mining, software engineering and HCI. He can be contacted at: sadaqat_jan@hotmail.com.