JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling

Olusola Babalola [1,*], Bolanle Ojokoh [2], and Olutayo Boyinbode [3]

[1] Department of Computer Science and Mathematics, Elizade University, Wuraola Ade.Ojo Avenue, P.M.B. 002, Ilara -Mokin, Ondo State, Nigeria; e-mail: olusola.babalola@elizadeuniversity.edu.ng
[2] Department of Information Systems, Federal University of Technology, P.M.B. 704 Akure, Ondo State, Nigeria; e-mail: baojokoh@futa.edu.ng
[3] Department of Information Technology, Federal University of Technology, P.M.B. 704 Akure, Ondo State, Nigeria; e-mail: okboyinbode@futa.edu.ng
* Corresponding Author : Olusola Babalola

Abstract: Topic modeling is an integral text mining component, employing diverse algorithms to uncover hidden themes within texts. This study examines the comparative performance of prominent topic modeling techniques on news headlines, which is characterized by brevity and specific linguistic style. Given the corpus originates from a non-native English-speaking country, an additional layer of complexity is introduced to the task. Our research explores the feasibility of employing a committee approach for topic modeling, evaluating the efficacy and challenges of various methods in practical settings. We applied three techniques—Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic—to create models with a fixed number of topics (n=40). These models were then tested on approximately 150,000 news headlines. To assess topic coherence, we utilized Word2Vec, human evaluators, and two large language models. Statistical tests confirmed the significance and impact of our findings. BERTopic demonstrated superior coherence compared to NMF, though slightly, but consistently outperformed NMF and LDA according to human and LLM evaluations. The notable disparity in LDA's performance relative to BERTopic and NMF underscores the importance of carefully selecting a topic modeling technique, as the choice can significantly influence the outcome of the analysis.

Keywords: Coherence Evaluation; Model Comparison; News Headlines; Non-Native English; Topic Modeling.

Received: October, 25th 2024
Revised: November, 6th 2024
Accepted: November, 12th 2024
Published: November, 23rd 2024

1. Introduction

Human communication mostly occurs online, and the rapid growth of online content has led to increasing demand for effective topic-modeling techniques that can handle short texts such as social media posts, text messages, and news headlines. Topic modeling is an unsupervised machine-learning process often used to group documents in a corpus into themes discernible from the corpus[1]. The algorithm infers themes from the documents and predicts the most applicable topic [2]. Topic modeling has great utility, particularly as it is an unsupervised learning process with various algorithms and techniques available for this task; essentially, the process considers documents to be a mixture of topics, each topic being a probability distribution over words[3]. A document is related to a range of topics, each organized around particular words. Topic models analyze the words found in documents in a corpus, identifying clusters of related documents, each cluster representing a distinct topic[4]. Abstract themes in the documents are found as clusters[3].

There are different approaches to discovering inherent themes in document collections using topic modeling, and some categories of topic modeling techniques include algebraic, fuzzy, probabilistic, and neural. Latent Semantic Annotation (LSA) and Non-negative Matrix Factorization (NMF) are well-known algebraic techniques. Latent Dirichlet Allocation (LDA) is a popular model of the probabilistic category, while BERTopic and LDA2Vec are examples of models from the neural models category[5]. Numerous studies have explored available

topic models, including[5]–[8]. Interested readers can explore topic models in depth from the study [9].

Probabilistic approaches model topics as probability distributions over words in a vocabulary based on the idea of a probabilistic generative process of documents. Matrix factorization techniques view topics as low-dimensional representations of the word-document matrix, decomposing the matrix into lower-dimensional matrices. While probabilistic approaches are based on theories that words in a document are generated from a probabilistic process and handle topic discovery is by reverse engineering this process, matrix factorization's target is to determine the low-dimensional representation of the original matrix that minimizes the reconstruction error. BERT-based embeddings (BERTopic) leverage contextual word embeddings and transformer architectures with the assumption that words in a document have different meanings depending on the context in which they appear. This enables the capture of semantic relationships in text through context rather than word co-occurrence statistics, addressing the limitations of probabilistic and matrix factorization methods.

As topic modeling algorithms became more utilized, their limitations began to surface. A limitation widely noted in the field is the performance of topic modeling algorithms over short texts[10]. Short texts primarily come from social media content or microblogging, where the platform restricts communication to a certain number of characters per document. The content space for text on social media sites like Twitter (now X) brought about an explosion of short text big data, necessitating the need for systems able to manage topic modeling of this sort of data. Short text is not only found in social media; other data sources with short-length combinations of characters include web page snippets, question-answer pairs, and status updates. News headlines also belong to this category. Applications analyzing this sort of dataset thematically, therefore, often utilize text topic modeling, especially when unable to go the supervised learning technique routes. Some characteristics of short text are a lack of enough co-occurrence information, the situation where text is probably generated by only one topic, and the statistical information of words among texts cannot capture semantically related words [11], [12]. Traditional topic models experience degradation over short text because of the poor co-occurrence of words in short text[11], [12].

Due to the challenges of short text topic modeling, the machine learning community has given attention to this problem and solutions reported in the literature, e.g.[10], [13], [14]. Despite research into specialist algorithms for short text, traditional topic modeling techniques are still popular among users. Examining recent surveys of topic modeling shows limited references and emphasis on the specialized topic modeling algorithms for short text [5], [15]–[17] or no references to it at all e.g. [9]. Of the survey literature mentioned, [17] devoted a section to "Short Text Optimized Topic Models". This situation suggests that the research community may continue to use generalized algorithms for short-text topic modeling tasks until specialized algorithms gain wider acceptance. Despite the low coverage in some literature, researchers are actively evaluating topic model performance on short texts with an interest in benchmarking techniques, like it is done for the generalized topic modeling algorithms. Common variations include varying the combination of topic models investigated or investigating the topic models with different datasets [10], [13], [14]. A distinct observation that most researchers use topic modeling in a suboptimal fashion was also raised, and appropriate recommendations were made in that regard[13].

Why another research on topic modeling performance and why short texts? This research focuses on the performance of topic modeling techniques applied to a unique dataset: a news headline corpus in English from a non-native English-speaking country. The dataset represents the "shorter end" of the short text category, with an average document length of only seven words. One objective of this research is to assess the effectiveness of established topic modeling methods in analyzing news headlines, specifically evaluating how much the choice of algorithm impacts the results. This study aims to compare newer topic modeling approaches against some classical approaches on a dataset of short news headlines. Specifically, we aim to compare performance extents regarding topic coherence, interpretability, and overall quality on the dataset, thus providing insights into the most effective approach for topic modeling on short texts.

Given the availability of several mature, state-of-the-art topic models, there is potential for combining these models to enhance topic prediction accuracy. The idea is to leverage a committee approach, where multiple models collaborate to determine the best fit for a document's topic. This research is a key step in exploring whether such a committee approach can

lead to improved topic predictions. By examining the performance of three different topic models, this study provides an empirical foundation to address whether utilizing a collaborative model framework can yield better outcomes in topic modeling for short texts like news headlines.

Our study aims to empirically evaluate the theoretical advantages and limitations of topic modeling approaches through systematic comparison of BERTopic, LDA, and NMF on short text datasets. We assess these models across multiple dimensions, including topic coherence and interpretability, considering the extent of their performance differences. This comparative analysis is particularly relevant given the increasing prevalence of short-form electronic content, where effective topic modeling can enhance various applications from content recommendation to trend analysis. Our research contributes substantial empirical evidence through systematic evaluation of topic coherence scores, LLM scoring of the topics generated, and human evaluation of these topics.

The rest of this paper is organized as follows. We examine the research literature in Section 2, reviewing related work and presenting key findings from the literature. Section 3 details our proposed method, including data acquisition, preprocessing, modeling techniques, and evaluation metrics. Specifically, Section 3.3 elaborates on the topic modeling algorithms employed, their parameters, and implementation details for LDA, BERTopic, and NMF. Section 4 presents the results and discussion, encompassing topic coherence analysis, statistical significance tests, descriptive statistics, non-parametric and permutation tests, MMW test, effect sizes, large language model evaluation, and human scoring of topics. Finally, Section 5 concludes the paper by summarizing our key findings and discussing potential future research directions.

2. Literature Review

A comparative study of methods for topic modeling in news[18] found the efficacy of Top2Vec and BERTopic over traditional topic modeling methods; the authors applied Top2Vec, BERTopic, NMF, LDA, and LSA to categorize a dataset of over 210,000 news headlines and abstracts. In another practical and empirical comparison of topic modeling models. Study [19] compared Top2Vec and the conventional (traditional) approaches of LDA and LSA on a dataset of over 65K COVID-19 abstracts. A key finding was that LDA and Top2Vec were highly correlated, followed by LSA and LDA.

A study of topic modeling-related scholarly articles between 2015 and 2020[10] examined commonly used topic modeling methods in text mining, testing LDA, LSA, NMF, PCA, and RP on short text. The results showed that LDA and NMF performed better than the others and were more consistent. Furthermore, the authors observed that fewer keywords led to a higher coherence score in LDA and NMF. The methods' data set was the 20 Newsgroup data set (20,000 documents) and the Facebook conversation data set (877 sentences and 7250 words).

In a similar research trend[10], five frequently used topic modeling methods used to analyze short-textual social data - LSA, LDA, NMF, Random Projection (RP), and Principal Component Analysis (PCA) - were evaluated using two datasets. LDA and NMF methods produced more meaningful topics and obtained good results based on topic quality and a few other metrics. A larger social media dataset was used by [8] to evaluate topic models; LDA, NMF, BERTopic, and Top2Vec were used to generate topic models for 31,800 unique tweets. The results were human-evaluated; BERTopic and NMF results were potent performers, followed by Top2Vec and LDA.

Using the 20Newsgroups dataset and a subset of the Yahoo Q&A (87,362 documents) to compare NMF, LDA, Paragraph Vector Topic Model (PVTM), Top2Vec and BERTopic topic modeling performance. Study [20] reported that Top2Vec and BERTopic performed prominently, with Top2Vec as the best performing across all datasets, followed by BERTopic and PVTM - all embedding techniques, outperforming LDA and NMF.

In an application for identifying the core themes in datasets consisting of tweets about diabetes and the reach within the Indian Twitter community [21], NMF outperformed LDA while BERTopic performed better than Top2Vec.

Turkish news was also used to compare topic modeling approaches[22]; LDA, LSA, NMF, and n-LDA were compared over a 4,200 news titles dataset. The accuracy was

compared, and NMF was found to be most successful for 3 classes, while LSA was the most successful for 5 and 7 classes. There were 7 class labels in the used dataset.

Similarly, for non-English news, the evaluation of topic models on Swedish news comparing NMF and LDA was undertaken [23]. The authors argued that the data on which topic models are learned is a main determinant of their usefulness. They observed that nouns yield the most meaningful topics in the case of Swedish news articles; however, proper nouns have the potential for misleading topics due to misclassification issues, leaving only common nouns as a solid choice for categorization. Their comparison of the topic models from NMF and LDA concluded that both had strengths and weaknesses depending on the specific use case.

2.1. Findings from the Literature

The prevalence of LDA was mentioned by [17] with the submission that many researchers are using this method despite it being a poor choice for modeling more complex data relationships that may occur in the textual data. LDA's popularity is not without scientific backing for its performance. Research [24] motivated LDA's usefulness for journalistic research and provided information about choices that influence the good performance of LDA (a sort of best practices information). While other topic modeling approaches existed, the early era largely established Latent Dirichlet Allocation (LDA) as a leading technique due to its effectiveness and versatility in topic discovery. Decades after, and even with more knowledge and developments in the field, stock LDA is as popular as before among topic modeling users. Figure 1 shows a chart of Google Trends data for LDA in the last five years. The chart of search volume, which may indicate popularity, shows a stable volume for LDA over the last five years. Given the availability of several other topic modeling options and newer approaches over the years, one inference may be that interest in LDA is not waning.

Other researchers, such as [7] have similarly raised the observation that a key cause of performance disparity among topic models in different settings is due to the variety of document types encountered by the topic models; hence, a topic model that performs excellently in one application may be woeful in another, especially when the dataset changes.

Determining what topic modeling method to use is strongly dependent on the data's properties and even the dataset's size. The dilemma sometimes is which topic modeling method to use despite the plethora of methods for undertaking the task and the impossibility of declaring one approach as the best overall choice further necessitates the current research work.

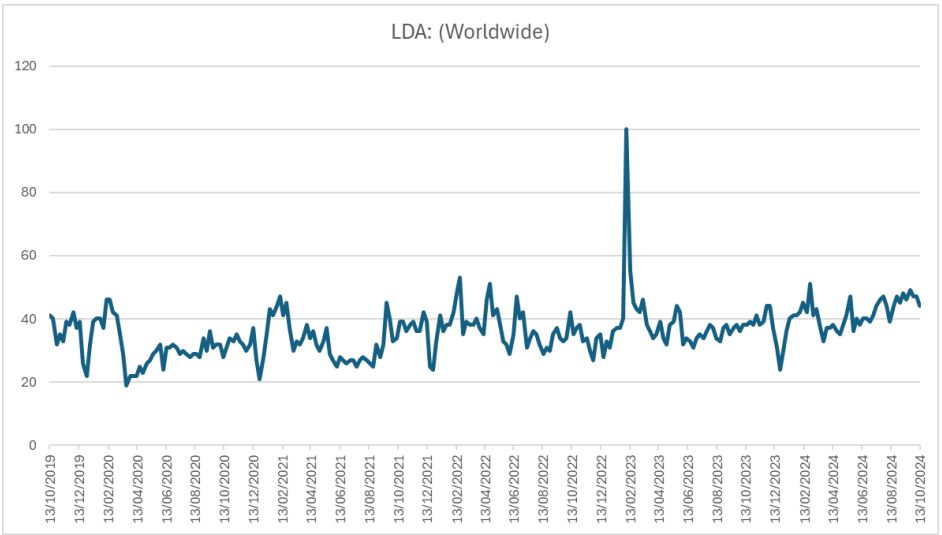

Figure 1. Chart of LDA searches from Google Trends data obtained October 2024.

3. Proposed Method

3.1. Data acquisition

The dataset was obtained by crawling a national newspaper website. This crawl was conducted in 2018 by one of the authors. The newspaper is completely in the English language,

although the country is a non-native English nation. The news coverage is national but includes key international news, especially in the politics and sports category. The Python programming language library, Scrapy, was used to build a web crawler, while the BeautifulSoup Python library was used to parse the webpages collected from the site. Robots.txt rules were respected. The webpages also contained advertisements as well as links to related content, but these were ignored, and the extraction process targeted HTML tags for the container holding the main news headline. In total, 149,679 articles were crawled and processed. The statistics are found in Table 1.

Table 1. Descriptive statistics of the corpus.

Statistic	Value
Count	149679 sentences
Mean	6.45 tokens
Standard Deviation	1.69 tokens
Minimum	1.0 tokens
25th Percentile (Q1)	5.0 tokens
Median (Q2)	6.0 tokens
75th Percentile (Q3)	7.0 tokens
Maximum	24.0 tokens

3.2. Preprocessing

Standard natural language processing (NLP) preprocessing pipeline tasks were undertaken, including stopword removal and tokenization using the NLTK library. The resulting preprocessed dataset has the characteristics listed in Table 1 concerning sentence length. The resulting preprocessed dataset had 149,679 different headlines. On average, headlines had about 6.45. Token lengths varied overall - some are very short, with only one word after preprocessing, while others are much longer, with 24 words. The standard deviation was 1.69, showing the sentence lengths are quite different. 25% of the sentences have five (5) words or less, and 75% have seven (7) words or less.

3.2. Modeling

3.3.1. Topic Modeling Algorithms Selection

One algorithm from each of the categories of topic modeling described in [5] was selected. From the algebraic class of topic models, NMF was selected; from probabilistic topic models, LDA was selected, and from the neural embedding class, the BERTopic model was selected. Each selected topic model has been widely tested in literature and is a relatively mature technique.

3.3.2. Model Parameters

Parameter settings can significantly impact the outcome of the topic modeling process, and therefore, careful consideration should be given to these. Knowledge of the dataset characteristics may help arrive at good parameter values. In Latent Dirichlet Allocation (LDA), two important parameters α and β control the document-topic density. A higher α value indicates a document is likely to contain a mix of most topics, while a lower value indicates a mix of few topics. This value typically ranges from 0.1 to 1.0 in most usage scenarios. The β parameter represents the topic-word density; a higher β value indicates topics containing most words while a lower β indicates topics are likely a mix of few words. When utilizing BERTopic analysis with HDBSCAN clustering, there are two crucial parameters to consider - minimum cluster size and minimum samples. In the context of BERTopic, a higher value of minimum cluster size constrains the algorithm from modeling only large clusters as topics, and a lower value considers clusters of smaller sizes as well. The minimum sample parameter determines what quantity of samples would form a dense region. A higher value means only denser regions will be considered as topics. One important parameter for Non-negative Matrix Factorization (NMF) is the regularization setting, which affects overfitting. A higher value means a more regularized model, which lowers the risk of overfitting. Topic modeling on short text presents unique challenges and requirements compared to traditional document-length text analysis.

All the three topic modeling approaches have implementations in Python programming language. We utilized LDA in the Gensim library, NMF in the scikit-learn library, and BERTopic from the bertopic package. The default values of the implementations were mostly maintained; except for the number of topics which we discuss later in this section, and the use of k-means clustering in BERTopic as against the default clustering approach. Appendix B contains the details about the key parameters for each of the tools, their defaults, and our own settings for those modified.

A necessary decision to be made during topic modeling is the number of abstract topics or themes to extract from the corpus. The number of topics affects the outcome of the topic modeling process. Choosing a value for k that is too low relative to the potential latent topics present in a corpus will produce topics that are too general, and choosing a k that is too large produces limited topics. Literature provides methods for analytically determining the appropriate number of topics to be used as a parameter in the topic modeling algorithm.

Based on the nature of the corpus and our research objective, k was heuristically determined. Since a key point of the current research is to consider the feasibility of a committee approach to topic modeling, it is necessary to be consistent with a common value of k for all the topic modeling approaches. Based on knowledge about the data, an estimate can be made for the corpus's possible abstract themes. Generally, newspapers such as the New York Times or The Guardian typically categorize their news. The number of news categories varies across newspapers and is often more than 20, with topics ranging from politics and business to sports and entertainment; in some other cases, 70 categories could be found in the newspaper. We choose the value of k=40 for the topic modeling experiments. By making k=40, the aim is to capture a comprehensive yet manageable presentation of the thematic landscape within the national newspaper discourse board. This research considers only the headlines, and this peculiarity would likely reduce the number of latent topics present in the corpus compared to when the entire news article content is modeled.

A curated dataset of the Huffington Post news articles[25] supports our assertion of the value of k = 40. The dataset creators found 42 key categories after pruning off categories with less than 1,000 news articles. The number of topics to use is an interesting discussion in the literature. For instance, [26] experimented with five topic modeling approaches on the dataset of 120,000 articles and noted that the optimal number of topics varied greatly across different approaches.

3.3.3. LDA Topic Modeling

The LDA modeling process used the Gensim library. The Gensim library in the Python ecosystem provides tools for conducting various NLP tasks, including topic modeling, word embedding, document indexing, and similarity retrieval. This library is used alongside re and NLTK packaged to build the LDA model. The LDA implementation in Gensim is based on [27], and a simplified algorithm for the process is presented in Listing 1. The preprocessing pipeline includes tokenization, conversion of case, punctuation removal, and stopword removal. The tokenized news headline is passed word by word into a function to create a Gensim dictionary that contains every unique token mapped to a unique integer. A special corpus is created by converting each news headline into a bag of words representation using the created dictionary. The LDA model is then trained on the corpus with the specified number of topics (40). Once the model is built, it is used to obtain the most probable topic of each news headline based on the distribution of topics learned. The top keywords that define each topic are extracted from the model. Ten (10) prominent words defining each topic were extracted. These are available in the supplementary data section of this article.

Algorithm 1. LDA Topic Modeling and Word Extraction Process

INPUT: List of headlines (document titles)
OUTPUT: List of topics, list of top words and topic words

```
Initialize list of stop words
For each document title:
    Tokenize the title into words
    Remove stop words from the tokenized title
Create a dictionary from the processed titles
Create a corpus from the processed titles using the dictionary
Train an LDA model using the corpus, dictionary, and desired number of topics
```

```
Initialize empty lists for topic labels, top words, and document-specific top words
For each topic in the LDA model:
    Get the top words for the topic
For each document in the corpus:
    Find the most probable topic for the document
    Get the top words for this topic
    Find words that appear both in the document and in the topic's top words
    Append the topic label, top words, document-specific words to respective lists
Return the lists of topic labels, top words, and document-specific words
```

3.3.4. BERTopic Topic Modeling

BERTopic is an advanced topic modeling approach leveraging Bidirectional Encoder Representations from Transformers (BERT), this enables it to achieve better semantic extractions from text collections. The use of BERT embeddings enables better capture of semantics. BERT transformers are used alongside class-based term frequency and inverse document frequency (c-TF-IDF) to create dense clusters of document contents. An overview of BERTopic is available in [28]. An official Python implementation for this topic modeling algorithm was available and used in this experiment with the steps presented in Listing 2. The preprocessing pipeline for building a BERTopic model is like the one described for the LDA. Similarly to the previous topic model, the number of topics was 40.

BERTopic modeling returns a topic with label -1 for for outliers. This topic comprises documents for which it cannot find a topic fit. Since the other models we are comparing do not possess this functionality, and since we are not interested in outliers because every content in the newspaper belongs to a discussion theme, no matter how remote their connection may be, this nature had to be adjusted. Potential workarounds to ensure every document is grouped were examined. BERTopic has a reduce-outliers function, for example, and the availability of Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering support could assist in a way. However, those options were not going to present the exact behavior desired. One way to ensure a methodology like the other two models' in ensuring 40 topics are generated with all documents getting assigned to one of those topics was to use HDBSCAN in combination with k-Means. This approach results in a model that does not have any outliers at all.

Algorithm 2. BERTopic Topic Modeling and Word Extraction Process

```
INPUT: List of headlines (document titles)
OUTPUT: List of topics, list of top words for each topic, list of topic assignments
Initialize list of stop words
For each document title:
    Remove stop words from the title
Train a BERTopic model on the processed titles, using HDBSCAN and K-Means clustering:
    Initialize a K-Means model with 40 clusters
    Pass the K-Means model as a parameter when initializing the BERTopic model
    Fit the BERTopic model to the preprocessed data
Get topics and probabilities for each document
Create a TF-IDF vectorizer and fit it to the processed titles
For each document:
    Get the assigned topic
    Find the most important words based on TF-IDF scores
    Get the top words for the assigned topic from the BERTopic model
    Store the topic, important words, and top words
Return the lists of topics, important words, and top words
```

A k-Means model forces every point of the BERTopic process to be fitted in a cluster. The k-Means model was initialized with 40 clusters, and the model was passed to the HDBSCAN model as a parameter when initializing the BERTopic model. The model is then fitted to the preprocessed data. As with the other cases, the top ten defining words from each of

the 40 topics of the trained model are extracted. These words are available in the supplementary information to this article.

3.3.5. NMF Topic Modeling

NMF is an unsupervised learning algorithm used for feature reduction and has been applied to text mining. The NMF topic modeling workflow used in this experiment is from the Scikit-learn library in Python. The NMF topic modeling approach analyses the data using NMF and TF-IDF techniques; a simplified process algorithm is presented in Listing 3. After the standard preprocessing operations, a TF-IDF vectorizer is created, and the vectorizer converts the news headline corpus into a matrix of TF-IDF features. The key parameters set are the cutoff of common terms and the minimum number of occurrences a word must have in the corpus to be considered. We ignore terms appearing in more than 95% of documents as well as those appearing in just one document. The vectorizer is then fitted to the corpus, and the NMF model is initialized to 40 topics; the random_state parameter is set to 42 to ensure consistent results across runs. The model is then fitted to the TF-IDF matrix. The top words defining each topic learned are extracted from the NMF generated topic models. These words are available as supplementary information to this article.

Algorithm 3. NMF Topic Modeling and Word Extraction Process

INPUT: List of headlines (document titles)
OUTPUT: List of topics, list of influential words for document, list of top words for topic

```
Initialize BERT model for sentence embeddings
Initialize list of stop words
For each document title:
    Tokenize the title into words
    Remove stop words from the tokenized title
    Join processed words back into a string
Create sentence embeddings for processed titles using BERT model
Initialize HDBSCAN for clustering
Initialize KMeans for restricting to 40 topics
Perform initial clustering using HDBSCAN
Apply KMeans to HDBSCAN clusters to restrict 40 topics
Initialize BERTopic model with custom clustering algorithms (HDBSCAN, KMeans)
Fit BERTopic model on processed titles and their embeddings
Initialize empty lists for topic labels, top words, and document topic assignments
For each unique topic in BERTopic model:
    Get the top words and their scores for the topic
    Append topic label and top words to respective lists
For each document title:
    Get the assigned topic from BERTopic model
    Append the topic assignment to document topic assignments list
Return the lists of topics, important words, and top words
```

3.4. Evaluation – Topic Coherence Computation

Topic coherence is a measure that provides an empirical examination of topics generated by a topic modeling algorithm and is useful for comparing the quality and the interpretability of those topics. The measure assesses the extent of semantic similarity between words in the same topic. The topic coherence score for each topic generated by the topic modeling algorithms under examination is computed. Word2Vec, a word embedding system that captures the context of words in a corpus, is used to compute this metric. The Word2Vec model is trained on the corpus and used to compute the similarity between the defining words of each topic. Utilizing Word2Vec introduces semantics as a main point in the computation of coherence. Research shows that the utility of word embedding-based metrics aligns with human preferences, robustly capturing the coherence of tweet topics [29]. Other coherence measures, including C_v, UMass are other popular measures that are also used in the literature.

The Word2Vec coherence computation method is based on the pairwise similarity of topic words in the vector space learned by the Word2Vec model. The output is the

coherence_score where each value represents the Word2Vec-based coherence score for the corresponding topic.

4. Results and Discussion

4.1. Topic Coherence Results

The topic coherence scores from the computation of Listing four are presented in Table 2. A set of violin plots to visualize the scores is in Figure 1. The tip and tail of each violin identify the highest coherence and the lowest coherence, respectively, for each of the topic models. The bulge of the body of the violin represents where the points are mostly clustered. The violin plots show that LDA coherence has the lowest tip and tail, NMF has a higher tip and higher tail than LDA, showing that the coherence scores are better, and BERTopic has the highest tip.

Table 2. Topic coherence scores.

Topic ID	LDA	NMF	BERTopic
1	0.596127	0.591615	0.652495
2	0.439613	0.8076	0.838678
3	0.540452	0.690843	0.746074
4	0.632748	0.706873	0.771809
5	0.55643	0.630687	0.726931
6	0.613984	0.700945	0.663604
7	0.670156	0.755528	0.728375
8	0.560395	0.708561	0.715568
9	0.539196	0.702046	0.797426
10	0.60417	0.693104	0.845845
11	0.520971	0.838268	0.769649
12	0.484266	0.60184	0.774371
13	0.5657	0.740172	0.78593
14	0.430545	0.783369	0.753339
15	0.422487	0.674193	0.767238
16	0.598895	0.758146	0.843414
17	0.54311	0.742988	0.863942
18	0.637431	0.603108	0.776688
19	0.541759	0.588079	0.637236
20	0.611433	0.68503	0.734866
21	0.545709	0.672195	0.467535
22	0.638131	0.845303	0.825441
23	0.480769	0.802615	0.850339
24	0.626074	0.62316	0.825968
25	0.462824	0.629499	0.71108
26	0.645126	0.6858	0.856945
27	0.524836	0.808476	0.757701
28	0.560657	0.768419	0.711628
29	0.507704	0.848924	0.890679
30	0.568419	0.801933	0.778918
31	0.665324	0.876758	0.841463
32	0.590228	0.776589	0.712149
33	0.492814	0.565418	0.758206
34	0.57503	0.660879	0.78806
35	0.516311	0.846189	0.635655
36	0.576584	0.774509	0.945186
37	0.552563	0.698167	0.680499

Topic ID	LDA	NMF	BERTopic
38	0.512765	0.753042	0.898287
39	0.510105	0.802577	0.598075
40	0.650618	0.743589	0.799372

4.1.1. Statistical Significance of Topic Coherence Results

This section analyzes the performance differences between Systems L(LDA), N(NMF), and B(BERTopic). The coherence result sets for the models are compared as one whole i.e. system-wide for the topic models, since the metrics across rows are unrelated. Therefore, we use descriptive statistics, non-parametric, and permutation tests to draw robust conclusions.

4.1.2. Significance of Differences

To establish whether differences exist between the results from L, N, and B results the Kruskal Wallis test is used. This test determines whether there are statistically significant differences between the medians of multiple independent groups. The Kruskal-Wallis and permutation tests provide convincing evidence to suggest significant differences in the performance distributions of the three systems. The Kruskal-Wallis test, a non-parametric equivalent of one-way ANOVA, yielded a highly significant p-value ($p < 0.001$), rejecting the null hypothesis of equal distributions across the three systems. Similarly, all pairwise permutation tests comparing the means of the systems resulted in p-values below the 0.05 significance level, further supporting the presence of statistically significant differences.

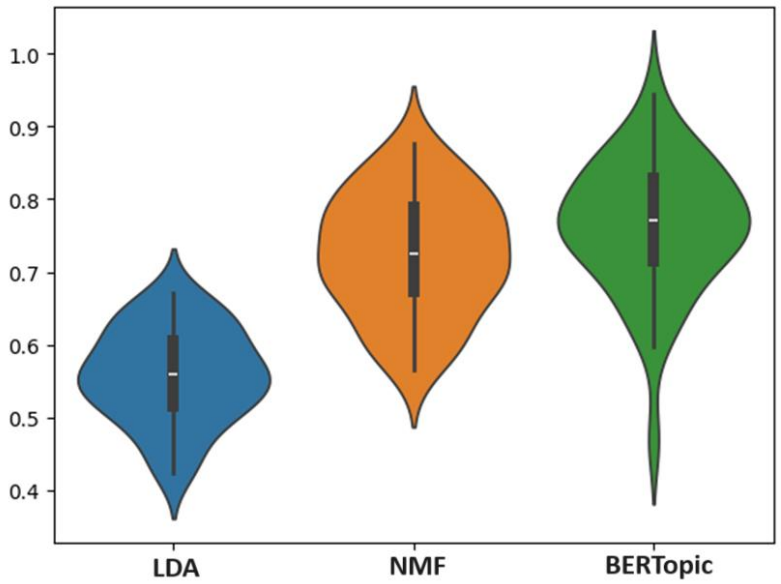

Figure 2. Visualization of distributions of Topic Coherence performance scores using violin plot

The test result is presented in Table 3, supporting the hypothesis that there are statistically significant differences in the topic coherence scores. The test result indicates a significant difference between the topic models' performance; therefore, we proceed to examine the difference in depth.

Table 3. Significance Tests Results.

Test	Statistic	P-value	Conclusion
Kruskal-Wallis	67.003	<0.001	Significant difference exists

4.2. Descriptive Statistics of Coherence Results

This section uses descriptive statistics to analyze the performance differences between Systems LDA, NMF, and BERTopic. As previously established, the data structure

necessitates methods that do not rely on independence assumptions between observations across rows.

Table 4 presents the descriptive statistics for each system's performance data. System B exhibits the highest mean performance (M = 0.763), followed by System N (M = 0.725) and then System L (M = 0.558). System B also demonstrates the largest variability (SD = 0.091) compared to the other two.

Table 4. Topic Coherence data descriptive statistics for LDA, NMF, and BERT coherence results.

Statistics	LDA	NMF	BERTopic
count	40.00000	40.00000	40.00000
mean	0.557811	0.724676	0.763167
std	0.064462	0.082258	0.090753
min	0.422487	0.565418	0.467535
25%	0.515424	0.673694	0.714713
50%	0.558412	0.724367	0.770729
75%	0.605986	0.788010	0.829146
max	0.670156	0.876758	0.945186

Table 4 provides a summary of the count, mean, standard deviation (std), minimum (min), lower quartile (25%), median (50%), upper quartile (75%), and maximum (max) for each of the column L, N, and B.

Considering overall performance, the mean values show System B exhibits the highest average performance (M = 0.763), which indicates better outcomes than Systems N (M = 0.725) and L (M = 0.558). The difference in the mean value between System L and Systems N and B is significant, suggesting a potentially substantial performance gap. Furthermore, System B has the greatest variability in performance (SD = 0.091), having a wider range between performances compared to System N (SD = 0.082) and System L (SD = 0.064). This result indicates that while System B may have higher performance, it is also possible in some regards that it would have a poor performance. We will return to examine this observation in Section 5.

Based on the percentiles, the following conclusions may be made: System L has the lowest spread; 50% of its values fall between 0.515 and 0.606, showing a consistent, lower performance profile. System N has a wider spread than System L. Its middle 50% of values range from 0.674 to 0.788, indicating a greater potential for both better and worse performance compared to System L. System B has the widest spread of all, with 50% of its values between 0.715 and 0.829 indicating the potential for the highest performance but also the greatest chances of variation.

4.3. Non-Parametric and Permutation Test

Pairwise Mann-Whitney-Wilcoxon (MWW) tests with Bonferroni correction were conducted to pinpoint specific pairwise differences after the significant Kruskal-Wallis result. All system pairs (L vs. N, L vs. B, N vs. B) showed statistically significant differences ($p < 0.05$) after correcting for multiple comparisons (Table 5). The observed differences and p-values indicate significant differences between L and N, and L and B, with a borderline p-value for the difference between N and B.

The permutation tests provide further support to the pairwise comparisons. These tests consistently show that System L has a significantly lower mean performance than Systems N and B. It is interesting to note the case between Systems N and B. While the mean difference between Systems N and B is smaller, it is still statistically significant at the 0.05 level according to the permutation test.

Table 5. Pairwise Permutation Tests Results.

Comparison	Observed Difference (Mean)	P-value
L vs. N	-0.167	<0.001
L vs. B	-0.205	<0.001
N vs. B	-0.038	0.050

4.4. MWW Test with Bonferroni Correction

The entire set of 40 coherence scores per model is compared as a single system. with that of the other models in a pairwise fashion using the Mann-Whitney-Wilcoxon (MWW) test. The MWW is a non-parametric statistical test that does not assume a particular data distribution. The test is conducted with Bonferroni correction to cater to multiple testing scenarios. The Bonferroni correction adjusts the significance level, and for each pairwise comparison, the corrected p-value is used to determine statistical significance. The corrected p-values address the increased chance of a Type I error because of multiple comparisons. If the corrected p-value is less than or equal to 0.05, the null hypothesis is rejected and a statistically significant difference exists between the groups. The null hypothesis holds if the corrected p-value is greater than 0.05, and any observed difference may be due to chance.

Since the comparison is between three systems, the MWW test is undertaken with multiple tests and corrections to adjust the significance level. The pairwise MWW test is between the coherence result sets for LDA versus MMF, LDA versus BERTopic, and NMF versus BERTopic. The complete data set is compared as a whole and not individual coherence for each topic; the results are presented in Table 6.

Table 6. MWW test for Topic Coherence data.

Test Type	Comparison	Statistic	P-value	Corrected P-value	Significant Difference
MWW	L vs. N	91.000	<0.001	<0.001	True
MWW	L vs. B	64.000	<0.001	<0.001	True
MWW	N vs. B	578.000	0.033	0.09917	False

The Mann-Whitney-Wilcoxon Tests show significant differences between groups L vs N, and L vs B, even after p-value correction.

The MWW tests with corrected p-values show that for L vs N and L vs B, the corrected p-values are much smaller than 0.05, and indicate statistically significant differences between them. The corrected p-value is greater than 0.05 for the N vs B pair, showing insufficient evidence to conclude a statistically significant difference between these groups. Statistical significance alone does not necessarily imply superiority in the practical sense; therefore, further examination is needed, considering the size of the effect and its relevance is one other step that is usually carried out for better understanding.

The results yielded statistically significant differences in key pairwise comparisons; between L and N, the U value and the corrected p-value suggest a substantial difference in the performance distributions of these two systems, with N demonstrating superior performance across the tasks. The p-value in this case shows that the observed difference cannot be attributed to chance and provides evidence from the results that N is better than L.

Similarly, comparing L and B yielded a favorable U value and a statistically significant corrected p-value. This result shows a marked difference in performance, with System B outperforming L. The extremely low p-value reinforces the statistical significance of this finding, indicating a clear difference in the systems' performance distributions.

The comparison between Systems N and B resulted in a Bonferroni-corrected p-value that is not statistically significant. While there may be a tendency towards B performing better, the evidence is insufficient to conclusively favor B over N in this pairwise comparison.

The results show that System B performs best, followed closely by System N; both systems have significant superiority over System L despite Bonferroni corrections. However, the comparison of Systems N and B fails the statistical significance test after the Bonferroni correction.

4.5. Effect Sizes

As earlier noted, the MWW tests alone may not be strong enough to establish conclusions, and the results can be examined to determine how much of an effect results from the test. Rank-Biserial Correlation is a desirable choice for effect sizes in the case of non-parametric tests like the Mann-Whitney-Wilcoxon test. It represents the difference between the proportion of favorable evidence for one group versus the other. The formula used for

calculating the effect size, specifically the rank-biserial correlation, for the Mann-Whitney-Wilcoxon test, is as in Equation (1).

$$r = 1 - \frac{2U}{n_1 n_2} \tag{1}$$

Where r is the rank-biserial correlation; U is the Mann-Whitney-Wilcoxon statistic; $n_1$and n_2 are the sample sizes of the two groups being compared.

This formula measures the effect size, which ranges from -1 to 1. A value of 0 indicates no effect, and values closer to -1 or 1 indicate a stronger effect. Positive values indicate that the first group has higher ranks on average, while negative values indicate that the second group has higher ranks on average.

The effect size is calculated to estimate the practical differences between the systems. U is the MWW statistics and n_1 and n_2 are the sample sizes of the two groups being considered. The effect size r provides a magnitude of the difference between n1 and n2, with values between 0.1, 0.3, and 0.5 typically considered small, medium, and large effects. Using the MWW test and accounting for multiple tests is a robust and conservative approach to comparing the performance of these three models. The effect sizes are also quantifiable. The effect sizes and their significance are found in Table 7.

Table 7. Mann-Whitney-Wilcoxon Test, including effect sizes.

Comparison	Statistic	P-value	Corrected P-value	Effect Size
L vs N	91.0	9.26e-12	2.78e-11	0.88625
L vs B	64.0	1.47e-12	4.41e-12	0.92
N vs B	578.0	0.03306	0.09917	0.2775

4.6. Large Language Model (LLM) Evaluation of Words from Topics

There are various measures to quantify the performance of topic models concerning quality, and topic coherence is just one of those. The literature covering this research field has shown that topic coherence scoring may not align with human coherence estimation. Rather than examine other traditional measures to assess the quality of topics, large language models are employed. The rise of chatbots and their increasing ability to perform knowledge-related tasks motivate this choice. The Meta AI based on Meta Llama 3.1, and the OpenAI GPT-4o-Mini chatbot based on GPT-4 were used to evaluate the topic models. The scoring system was basic; the bot was instructed to deduct one (1) for each word it considered not applicable to the theme but present in the set of defining words of the topic. The two LLMs were provided with similar prompts (Figures 3a and 3b) as well as the defining words for each topic model (Appendix A1-A3). The combined results of two LLMs for each of the topic modeling systems are presented in Table 8.

You will assume the role of an expert on Nigeria, utilize your knowledge from Nigeria news, social media and information on the Internet. Each row contains ten words. Each row should ideally form a topic of discussion, and all those 10 words should contribute to that same topic. For each word that you are not able to place in the topic along with others in the row, subtract 1 from the row score. The row score for a perfect row ideally would be 10. A row where all the words are not related and so no central topic can be inferred would score a 0.

(a)

Great. You should assign a score to each row. You will assume the role of an expert on Nigeria, utilize your knowledge from Nigeria news, social media and information on the Internet. Each row contains ten words. Each row should ideally form a topic of discussion, and all those 10 words should contribute to that same topic. For each word that you are not able to place in the topic along with others in the row, subtract 1 from the row score. The row score for a perfect row ideally would be 10. A row where all the words are not related and so no central topic can be inferred would score a 0.

(b)

Figure 3. (**a**)Prompt supplied to the first LLM along with top words of the topic models; (**b**) Prompt supplied to the second LLM along with top words of the topic models

Table 8. LLM scoring for topic models.

Topic	LDA		NMF		BERTopic	
	Meta AI	GPT-4	Meta AI	GPT-4	Meta AI	GPT-4
1	9	3	6	2	9	6
2	8	8	9	9	10	9
3	8	5	7	4	10	7
4	9	6	10	10	10	8
5	8	4	8	3	9	6
6	6	3	9	5	10	9
7	10	10	8	3	10	8
8	8	5	9	8	9	7
9	8	4	9	6	10	8
10	9	5	9	5	10	9
11	8	6	10	10	10	7
12	8	7	7	4	9	6
13	8	3	9	4	9	5
14	8	5	10	2	9	6
15	8	4	8	3	9	7
16	10	7	10	10	10	8
17	9	10	9	9	10	9
18	9	8	8	5	9	7
19	8	3	6	1	10	8
20	7	10	9	7	9	6
21	8	3	8	3	10	7
22	9	5	10	10	9	8
23	8	4	10	10	10	7
24	8	5	9	7	10	8
25	10	2	9	4	9	7
26	10	4	8	4	10	6
27	9	5	10	10	10	8
28	8	6	10	9	8	9
29	8	3	10	10	10	8
30	8	4	9	3	10	7
31	9	4	10	10	10	8
32	7	5	10	10	9	9
33	8	5	9	4	10	7
34	8	9	10	10	10	8
35	8	6	10	9	9	6
36	9	4	10	10	10	7
37	8	4	10	8	9	5
38	8	3	10	10	10	6
39	9	5	10	9	6	5
40	9	2	9	3	8	6
Total	**335**	**204**	**361**	**263**	**378**	**288**

The summary table comprising the totals is presented in Table 9.

Table 9. Summary for LLM scoring for topic models.

Topic Model	Total (Meta AI)	Total (ChatGPT-4)
LDA	335	204
NMF	361	263
BERTopic	378	288

4.7. Human Scoring of Topics

To incorporate a human scoring element in the analysis of the topic models' quality, the list of defining words for each topic was provided to a human evaluator familiar with the local community for assessment. The evaluator is a resident citizen, currently an undergraduate student, with a broad knowledge of general topics in society. Since there are ten words per topic, the evaluator was instructed to deduct one for every word judged as not belonging or not fitting with the others. This results in each topic receiving a score less than or equal to 10 for each of the topics generated by the models. As there was only one evaluator, there was no inter-rater reliability testing.

Table 10. Topic quality scoring by the human evaluator.

Topic	LDA	NMF	BERTopic
1	10	6	10
2	2	7	10
3	8	8	10
4	8	7	9
5	4	6	10
6	6	10	8
7	10	9	10
8	7	6	6
9	6	7	10
10	6	8	10
11	6	7	10
12	6	2	8
13	6	10	10
14	3	10	10
15	10	6	8
16	10	10	10
17	9	10	10
18	8	6	10
19	6	3	9
20	5	10	10
21	8	10	9
22	9	10	10
23	5	10	10
24	9	10	10
25	9	8	10
26	10	7	10
27	7	9	10
28	7	9	10
29	7	10	10
30	10	10	10
31	10	10	10
32	5	10	10
33	7	8	10
34	8	10	10
35	1	10	10
36	9	8	9
37	8	10	9
38	3	10	10
39	3	10	6
40	9	8	7
Total scores	**280**	**335**	**378**

Table 10 shows the breakdown of the scores by topic for each model, while Figure 4 shows a chart comparison of the scores.

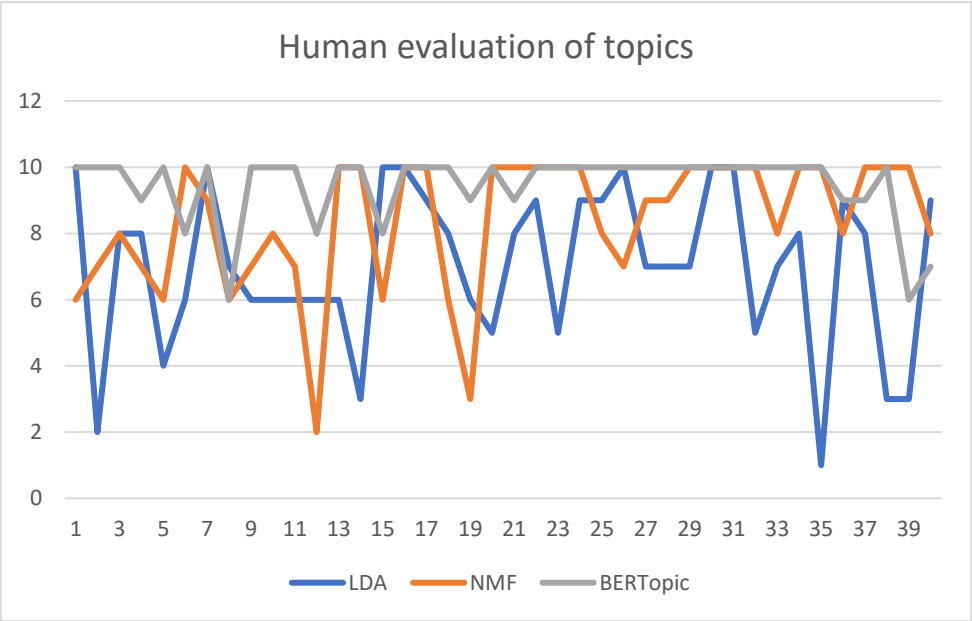

Figure 4. A line chart showing the human evaluator scoring of the three topic models.

5. Conclusions

The current research in the vein of comparative investigations of topic model performance on real-world data provides insights into topic modeling of news headlines in the form of short text. In this case, the performance of LDA, based on human evaluation and coherence measures, was below that of NMF and BERTopic. Researchers who may wish to use LDA topic modeling as a methodology on a similar dataset may wish to note this. Modifications to the LDA technique that work for short text could result in better performance; therefore, it is recommended that researchers consider those. Where there is a reason making it impossible to do so, then the use of other topic modeling algorithms investigated to work for short text is suggested.

Statistical tests of the results rate NMF better than BERTopic; however, human and LLM evaluations suggest otherwise. Although the p-value was below the corrected threshold, the effect size was relatively smaller, suggesting that the practical significance may be lower. The human evaluation of the predicted topics for a news headline favors BERTopic over NMF despite the statistical difference between NMF and BERTopic.

Statistical analysis allows a conclusion based on coherence values. However, the validation of the topic models by humans and both LLMs does not totally agree with the statistical tests. For example, examining the top words for topics shows there is a noticeable difference in quality between LDA and BERTopic models. However, the MWW test does not inform significance. However, a high effect size suggests notable practical differences between LDA and BERTopic.

Our findings have significant implications for practitioners in various domains. For applications requiring real-time processing or deployment on resource-constrained environments, NMF presents a compelling alternative to BERTopic, given their statistically comparable performance ($p = 0.033$). However, BERTopic's marginally superior coherence scores suggest it is the preferred choice for applications where semantic accuracy is key and computational resources are available. Users must weigh these trade-offs against their specific scalability, interpretability, and computational efficiency requirements when selecting a topic modeling approach for short text analysis.

One key information from the statistical results was that BERTopic was highly variated. A detailed examination of the coherence scoring shows that the lowest-performing topics of BERTopic were Topics 21, 35, 19, 1, and 6, with Topic 21 being an outright outlier with a value of 0.467535. Examining the scoring of this topic by the other evaluation processes

shows Topic 21 was scored 9 by the human evaluator and scored 10 and 7, respectively by the LLM evaluator. Without this outlier, BERTopic would have a more positive advantage from the data. BERTopic's superior performance can be attributed to its advanced handling of contextual understanding through pre-trained transformer architectures. With the ability to capture semantic relationships through contextual embeddings, BERTopic is highly advantageous in short text scenarios with limited word interplay, which may challenge techniques limited solely to word co-occurrence patterns.

The uniformity of the performance ranking of the three topic models in this experiment across the various evaluations established decent performance of BERTopic and NMF topic models, with LDA being outranked overall. The good performance of NMF in this experiment is worthy of note, as it supports results from previous literature that considered short-text datasets similar in structure to the one currently used [30].

The limitations of the methodology include the out-of-the-box usage of the various implementations of the techniques, especially as there may be tweaks leading to improved performance. Furthermore, we relied on coherence to make the comparison. The application in which we intend to utilize topic modeling requires that we have topics that are coherent to humans, hence the bias towards topic coherence when comparing their performance.

Author Contributions: Conceptualization: Sola Babalola; methodology, Sola Babalola and Bolanle Ojokoh.; software: Sola Babalola; validation: Bolanle Ojokoh and Olutayo Boyinbode. and Sola Babalola; formal analysis: nil.; investigation: Sola Babalola, Bolanle Ojokoh; resources: Bolanle Ojokoh and Olutayo Boyinbode; data curation: Sola Babalola; writing—original draft preparation: Sola Babalola; writing—review and editing: Sola Babalola, Bolanle Ojokoh, and Olutayo Boyinbode; visualization: Sola Babalola; supervision: Bolanle Ojokoh and Olutayo Boyinbode; project administration: Bolanle Ojokoh and Olutayo Boyinbode; funding acquisition: nil.

Funding: This research received no external funding.

Data Availability Statement: Due to copyright risks, the data used in this research is not made available. However we provide the top ten (10) words from the topics.

Acknowledgments: The authors acknowledge the support of Emmanuel Akinjogunla towards organizing some data aspects of the research.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Natl. Acad. Sci.*, vol. 101, no. suppl_1, pp. 5228–5235, Apr. 2004, doi: 10.1073/pnas.0307752101.
[3] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *Ann. Appl. Stat.*, vol. 1, no. 1, Jun. 2007, doi: 10.1214/07-AOAS114.
[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," in *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, Jul. 2004. [Online]. Available: http://arxiv.org/abs/1207.4169
[5] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.
[6] Y. Chen, Z. Peng, S.-H. Kim, and C. W. Choi, "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review," *Commun. Methods Meas.*, vol. 17, no. 2, pp. 111–130, Apr. 2023, doi: 10.1080/19312458.2023.2167965.
[7] R. Churchill and L. Singh, "The Evolution of Topic Modeling," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–35, Jan. 2022, doi: 10.1145/3507900.
[8] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.
[9] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *ICST Trans. Scalable Inf. Syst.*, p. 159623, Jul. 2018, doi: 10.4108/eai.13-7-2018.159623.
[10] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front. Artif. Intell.*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.
[11] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, Mar. 2022, doi: 10.1109/TKDE.2020.2992485.

[12] J. Qiang, P. Chen, T. Wang, and X. Wu, "Topic Modeling over Short Texts by Incorporating Word Embeddings," in *Advances in Knowledge Discovery and Data Mining*, 2017, pp. 363–374. doi: 10.1007/978-3-319-57529-2_29.
[13] C. D. P. Laureate, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *Artif. Intell. Rev.*, vol. 56, no. 12, pp. 14223–14255, Dec. 2023, doi: 10.1007/s10462-023-10471-x.
[14] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, "Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling," *IEEE Access*, vol. 10, pp. 105328–105351, 2022, doi: 10.1109/ACCESS.2022.3211396.
[15] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, "Topic Modelling Meets Deep Neural Networks: A Survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Aug. 2021, pp. 4713–4720. doi: 10.24963/ijcai.2021/638.
[16] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Appl. Sci.*, vol. 13, no. 2, p. 797, Jan. 2023, doi: 10.3390/app13020797.
[17] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.
[18] S. D. Rajan, T. Coombs, M. Jayabalan, and N. A. Ismail, "A Comparative Study of Methods for Topic Modelling in News Articles," in *Data Science and Emerging Technologies*, 2024, pp. 269–277. doi: 10.1007/978-981-97-0293-0_20.
[19] R. Thomson, E. Cranford, S. Somers, and C. Lebiere, "A Novel Approach to Intrusion Detection Using a Cognitively-Inspired Algorithm," in *Hawaii International Conference on System Sciences 2024*, 2024. doi: 10.24251/HICSS.2023.116.
[20] A. Amaro and F. Bacao, "Topic Modeling: A Consistent Framework for Comparative Studies," *Emerg. Sci. J.*, vol. 8, no. 1, pp. 125–139, Feb. 2024, doi: 10.28991/ESJ-2024-08-01-09.
[21] T. Ramamoorthy, V. Kulothungan, and B. Mappillairaju, "Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India," *Front. Artif. Intell.*, vol. 7, Feb. 2024, doi: 10.3389/frai.2024.1329185.
[22] Z. A. Güven, B. Diri, and T. Çakaloğlu, "Comparison of Topic Modeling Methods for Type Detection of Turkish News," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Sep. 2019, pp. 150–154. doi: 10.1109/UBMK.2019.8907050.
[23] J. Blad and K. Svensson, "Exploring NMF and LDA Topic Models of Swedish News Articles," Uppsala Universitet, 2020. [Online]. Available: https://uu.diva-portal.org/smash/get/diva2:1512130/FULLTEXT01.pdf
[24] C. Jacobi, W. van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journal.*, vol. 4, no. 1, pp. 89–106, Jan. 2016, doi: 10.1080/21670811.2015.1093271.
[25] R. Misra, "News Category Dataset," *arXiv*. 2022. [Online]. Available: https://arxiv.org/abs/2209.11429
[26] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, H. Qin, and X. Guo, "Search for K: Assessing Five Topic-Modeling Approaches to 120,000 Canadian Articles," in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 3640–3647. doi: 10.1109/BigData47090.2019.9006160.
[27] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for Latent Dirichlet Allocation," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 2010, pp. 856–864.
[28] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv*. Mar. 11, 2022. [Online]. Available: http://arxiv.org/abs/2203.05794
[29] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Jul. 2016, pp. 1057–1060. doi: 10.1145/2911451.2914729.
[30] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between LDA and NMF based Schemes," *Knowledge-Based Syst.*, vol. 163, pp. 1–13, Jan. 2019, doi: 10.1016/j.knosys.2018.08.011.

Appendix A

Table A1. Top words in LDA topics.

ID	LDA Topic	Coherence
1	man year old woman three police arrested allegedly arrest seven	0.576584
2	oil dies reps probe arrests top say award panel child	0.512765
3	ondo one cbn ibadan kidnappers lagos drug police ncc rate	0.507704
4	league strike must champions face without premier give leave results	0.54311
5	open report ready abia french saudi arraigns labour training records	0.645126
6	students power banks firms account trade moves non love rape	0.480769
7	pdp apc election inec osun says atiku wike buhari declares	0.60417
8	air kwara airport media tax monarch abuja force fans polls	0.598895
9	school corruption anti war baby elections pupils ibom plateau ig	0.665324
10	bank get business law wants economy firm plans promises threatens	0.439613
11	fire months real varsity student oyo madrid theft oshiomhole accident	0.626074
12	market african killing chairman boy funds inaugurates technology expert victory	0.638131

ID	LDA Topic	Coherence
13	assembly national win sector fund chelsea loan arsenal another bn	0.650618
14	world cup warns breaking wife back husband militants river france	0.539196
15	urges africa south gets fight life car meets security projects	0.57503
16	haram boko army military kills police suspects robbery suspect releases	0.5657
17	nigerians stop kano return presidential coach service buhari poor electricity	0.568419
18	kill attack police suspected herdsmen gunmen borno benue two ban	0.637431
19	bn fg children nigeria pay make nnpc loses tn states	0.484266
20	people lasg bid policeman protests like chinese head card cash	0.632748
21	bayelsa customs dss rice office unveils aide increase change last	0.520971
22	girls daughter church enugu governor pastor parents lawyer chibok local	0.422487
23	death years day naira end experts jailed liverpool member start	0.492814
24	crisis beat final ronaldo foreign united england vs speaker lead	0.524836
25	president buhari icymi eagles house presidency best says violence next	0.430545
26	buhari osinbajo photos state kaduna leaders meet killings economic fresh	0.540452
27	set ekiti fayose family hits russia financial hold go political	0.611433
28	edo updated dead anambra poll un imo gov missing team	0.670156
29	workers protest trump members china teachers suspends payment us hit	0.55643
30	killed ogun six die clash begins crash money injured two	0.560395
31	court ex efcc alleged fraud boss bn trial orders soldiers	0.590228
32	son city father men el home still man want officials	0.560657
33	rivers health leader know gives insurance things official review joshua	0.462824
34	chief wins release katsina west takes lament players tackle scam	0.552563
35	sex jonathan education girl campaign driver rejects marriage bail fifa	0.613984
36	residents community high job akeredolu appoints food commissioner lagos investors	0.596127
37	attacks illegal sign officers help makes victims signs medical navy	0.541758
38	budget mother time fake land building need frsc tell restructuring	0.516311
39	seeks group niger support delta north nigeria korea recession seek	0.510105
40	minister youths schools advises list th fct union ajimobi rise	0.545709

Table A2. Top words in NMF topics.

ID	NMF Topic	Coherence
1	nigerian youths women uk army economy students arrested india school	0.774509051
2	die crash auto ogun injured dead road accident dies ibadan	0.753042307
3	lagos ibadan expressway ambode residents airport arrested marathon pupils land	0.84892385
4	efcc fraud alleged arraigns probe arrests jonathan corruption trial case	0.742987691
5	ex gov minister boss chief deputy dies militants son speaker	0.685799507
6	bank account access cbn diamond robbery accounts skye loan sterling	0.802615324
7	president mr elect vice fifa zimbabwe updated senate visit sworn	0.693103984
8	photos osinbajo meeting abuja visits meets presides fec protest meet	0.758146043
9	buhari meets presidency aisha corruption congratulates obasanjo mourns tinubu anti	0.876758324
10	killed clash injured soldiers rivers feared cult suicide updated fresh	0.807599718
11	league premier champions english results table uefa latest things city	0.623160188
12	tells don stop group husband pay youths leaders leave obasanjo	0.845303151

ID	NMF Topic	Coherence
13	death toll rises son stabs sentences hits allegedly crushes mother	0.743589274
14	year old girl boy rapes raping woman daughter 10 rape	0.702046149
15	says ll won ve didn presidency osinbajo obasanjo ready soon	0.660879048
16	election inec ondo edo osun anambra gov poll candidate wins	0.740172176
17	attack herdsmen benue suicide killings kaduna dead fulani kills fresh	0.801933254
18	fg states power urges projects girls asuu plan chibok plans	0.603108204
19	000 10 years jobs 20 100 youths gets bail 30	0.60184016
20	budget assembly senate national 2018 2017 saraki reps probe 2016	0.706873378
21	south africa east west african north korea zuma leaders govs	0.838268488
22	police arrest suspected suspects robbery kidnappers nab recover killing robbers	0.674193333
23	world cup 2018 eagles russia fifa final england win france	0.565418071
24	seeks support govt ambode group urges security lasg women education	0.8084761
25	man city united allegedly utd mourinho stealing years jailed guardiola	0.783368707
26	new gets appoints unveils york signs deal coach introduces policy	0.690842865
27	apc primary gov rivers ondo members primaries tinubu oshiomhole chieftain	0.685029756
28	workers strike protest salaries health unpaid begin pay students asuu	0.755528144
29	pdp sheriff makarfi crisis convention chairman candidate chair members atiku	0.630687195
30	wife husband woman kills children sex son divorce pregnant marriage	0.708561102
31	oil nnpc market price firms gas opec prices production sector	0.776589312
32	court orders remands supreme appeal case jails suit trial bail	0.768418888
33	nigerians urges libya return million react uk health warns advises	0.629498917
34	2019 ll elections inec warns atiku polls win presidential presidency	0.698167388
35	kill gunmen rivers suspected abduct policemen kidnap robbers benue troops	0.700945192
36	nigeria recession needs economy restructuring economic investment trade ll uk	0.591615162
37	delta niger militants avengers youths leaders community arrests pipeline peace	0.588079202
38	haram boko borno kills army troops soldiers attacks members military	0.846188581
39	fayose ekiti fayemi gov poll govt assembly dss lawmaker primary	0.802577004
40	trump north korea clinton ban house russia obama putin white	0.672194876

Table A3. Top words in BERTopic topics.

ID	BERTopic Topic	Coherence
1	wife marriage husband my wedding pope court woman tells divorce	0.945185808
2	inec election ekiti fayose poll edo fayemi osun gov ondo	0.898286707
3	delta niger oil nscdc pipeline navy avengers ndelta militants arrests	0.890679138
4	efcc fraud court alleged trial bail scam jonathans case witness	0.86394241
5	bank fg market insurance banks tax recession nse debt profit	0.856944704
6	buhari buharis obasanjo tinubu says osinbajo presidency president tells meets	0.850339423
7	police arrest gunmen robbery kidnappers suspected kill two lagos robbers	0.845845184
8	osinbajo dies mourns olubadan photos ajimobi oba ooni mourn benin	0.843414382
9	joshua olympics nigeria win tourney athletes cup tennis boxing sports	0.841462735
10	league champions arsenal city ronaldo mourinho chelsea madrid man messi	0.838677985
11	students education school teachers schools jamb pupils varsity health utme	0.825968244
12	sex cancer study children health women expert mental breast surgery	0.825441485
13	man woman death allegedly rape girl baby daughter sex police	0.799371702

ID	BERTopic Topic	Coherence
14	trump us north korea trumps saudi clinton russia brexit president	0.797425525
15	mugabe zimbabwe president gambia zuma opposition election jammeh pilgrims south	0.788060228
16	herdsmen benue taraba fulani killings cattle ortom kill grazing zamfara	0.785929571
17	cbn naira forex dollar rate inflation market banks cbns reserves	0.778918031
18	business facebook how ways account digital media data smartphone mobile	0.776688064
19	nigeria fg nigerias trade investment economy economic recession lagos nigerian	0.774370501
20	rivers police kaduna nysc killings wike protest rerun river dss	0.771809047
21	nigeria ambode nigerians nigerias lagos nigerian says poverty restructuring urges	0.769648737
22	fire crash frsc die flood accident injured kills auto killed	0.76723804
23	hiv cholera meningitis malaria polio outbreak monkeypox hivaids sickle who	0.758206195
24	eagles cup world fifa super rohr nff afcon vs falcons	0.757701043
25	rice customs lagos road roads lagosibadan rail bridge fg water	0.753339118
26	apc pdp sheriff candidate primary convention gov atiku 2019 oshiomhole	0.746074469
27	workers strike budget salaries pension wage minimum unpaid fg salary	0.734865756
28	music bbnaija awards women award davido im my wins prize	0.728375055
29	oil power nnpc fuel electricity gas price petrol supply firms	0.72693132
30	ipob igbo kanu nnamdi anambra okorocha biafra restructuring okada ohanaeze	0.715568117
31	airport air airlines aviation flight airports ncaa flights arik abuja	0.712148564
32	attack libya suicide migrants killed bomb bombers idps nigerians south	0.711628249
33	boko haram chibok girls borno troops army dapchi kills bharam	0.711079619
34	senate court judges corruption saraki melaye njc supreme cct melayes	0.680498882
35	lasg seeks residents waste lasema workers urges buildings moves wants	0.663603793
36	open serena nadal federer djokovic wimbledon murray tennis venus williams	0.652495492
37	lassa fever kills latest table english confirms premier cases doctor	0.63723636
38	drug ndlea trafficking drugs cocaine nafdac arrests hemp human cannabis	0.635654775
39	eavesdropper collated english results premier league inexcusable wk it	0.598075062
40	ebola congo zika dr who outbreak virus cosby shopaholic vaccine	0.467535008

Appendix B

Table B1. Parameters used for topic modeling

Category	LDA	NMF	BERTopic
Python Library	Gensim	- sklearn (NMF)	- BERTopic
Key Parameters	- number of topics alpha, Dirichlet prior for document-topic distribution - beta/eta, Dirichlet prior for topic-word distribution - maximum iterations - inference method - random seed - number of jobs/threads	- number of topics/components - initialization method - optimization solver - loss function - maximum iterations - alpha (regularization parameter) - l1_ratio (regularization mixing parameter) - random seed	- number of topics - number of top words per topic - minimum topic size - n_gram_range - calculation method - language - diversity factor - transformer model

Category	LDA	NMF	BERTopic
Default Settings	- n_components: 10 - doc_topic_prior (alpha): 1/n_components - topic_word_prior (beta/eta): 1/n_components - max_iter: 10 - learning_method: 'online' - random_state: None - n_jobs: None (uses all available cores)	- n_components: 10 - init: 'nndsvd' - solver: 'cd' (Coordinate Descent) - beta_loss: 'frobenius' - max_iter: 200 - alpha: 0.0 - l1_ratio: 0.0 - random_state: None	- nr_topics: "auto" - top_n_words: 10 - min_topic_size: 10 - n_gram_range: (1, 1) (unigrams) - calculation_method: 'umap' - language: "english" - diversity: 0.5 - embedding_model: 'all-MiniLM-L6-v2'
Modified Parameters	n_components = 40	n_components = 40; random_state = 42	nr_topics = 40; k-means clustering;

.