Exploring the Effects of Preprocessing Techniques on Topic Modeling of an Arabic News Article DataSet

Haya Alangari * and Nahlah Algethami

Posted Date: 20 November 2024

doi: 10.20944/preprints202411.1509.v1

Keywords:

BERTopic; topic modeling; pre-processing techniques; LDA; NMF; NPMI; topic diversity; Tashaphyne; ISRI

 doi:10.20944/preprints202411.1509.v1

Article

Exploring the Effects of Pre-processing Techniques on Topic Modeling of an Arabic News Article Data Set

Haya Alangari and Nahlah Algethami *

College of Computing and Informatics, Saudi Electronic University, Riyadh, 1167 Saudi Arabia
* Correspondence: n.algethami@seu.edu.sa;

Abstract: This research investigates the impacts of pre-processing techniques on the effectiveness of topic modeling algorithms for Arabic texts, focusing on a comparison between BERTopic, Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF). Using the Single-label Arabic News Article Data set (SANAD), which includes 195,174 Arabic news articles, this study explores pre-processing methods such as cleaning, stemming, normalization, and stop word removal, which are crucial processes given the complex morphology of Arabic. Additionally, the influence of six different embedding models on the topic modeling performance was assessed. The originality of this work lies in addressing the lack of previous studies that optimize BERTopic through adjusting the *n*-gram range parameter and combining it with different embedding models for effective Arabic topic modeling. Pre-processing techniques were fine-tuned to improve data quality before applying BERTopic, LDA, and NMF, and the performance was assessed using metrics such as topic coherence and diversity. Coherence was measured using Normalized Pointwise Mutual Information (NPMI). The results show that the Tashaphyne stemmer significantly enhanced the performance of LDA and NMF. BERTopic, optimized with pre-processing and bi-grams, outperformed LDA and NMF in both coherence and diversity. The CAMeL-Lab/bert-base-arabic-camelbert-da embedding yielded the best results, emphasizing the importance of pre-processing in Arabic topic modeling.

Keywords: BERTopic; topic modeling; pre-processing techniques; LDA; NMF; NPMI; topic diversity; Tashaphyne; ISRI

1. Introduction

As the digital universe expands, so does the volume of unstructured text available in various languages, including Arabic, presenting challenges and opportunities for computational linguistics. The growth in digital Arabic text data further underscores the need for advanced computational approaches to manage and analyze such content effectively. This situation elevates the significance of automated methods for extracting information, collectively known as information retrieval methods [1]. The ability to autonomously discern and organize topics across this extensive array of text not only sheds light on content patterns and information retrieval strategies, but also enhances document categorization and other related areas. Topic modeling is a sophisticated concept in the field of text analysis, particularly within the domains of Natural Language Processing (NLP) and information retrieval. It involves the automatic identification and extraction of relevant themes or topics from a vast collection of documents. This process is essential for understanding the underlying thematic structure in large volumes of text data, often being used to organize, manage, and provide insights into unstructured text [2].

Topic modeling fundamentally relies on statistical techniques to uncover abstract topics within documents. It assumes that each document in a collection is a mixture of various topics, where each topic is characterized by a distribution of words. The goal of topic modeling algorithms is to reverse engineer this mixture and identify the topics that best represent the collection of documents [3].

Traditional models, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), have laid the groundwork for understanding text corpora through statistical methods that analyze the distribution of words across documents to uncover latent thematic structures [4,5]. However, the advent of deep learning and neural networks has paved the way for more sophisticated models

 doi:10.20944/preprints202411.1509.v1

that can grasp the nuances of language and context more effectively. These methods aim to address the limitations of traditional models, particularly in handling large and complex data sets, better capturing the nuances of language, and improving model interpretability. Some notable advancements include Dynamic Topic Models (DTMs), Non-Negative Matrix Factorization (NMF) with Advanced Regularization, neural topic models, and BERTopic [6,7].

This study makes a significant contribution to the field of Natural Language Processing (NLP) by providing an in-depth analysis and comparison of the BERTopic model with traditional topic modeling techniques, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), specifically within the context of Arabic text analysis. This research examines the impact of various pre-processing strategies, with a particular focus on enhancing the capacity of the BERTopic model to generate coherent and diverse topics from Arabic news articles. Furthermore, it explores hyperparameter tuning, investigating how different configurations influence BERTopic's performance in terms of coherence and diversity, with the goal of establishing optimized settings for Arabic text analysis. To the best of our knowledge, this is the first study to systematically optimize BERTopic by adjusting the *n*-gram range parameter and integrating different embedding models for effective Arabic topic modeling. The findings underscore the importance of a nuanced approach to pre-processing and hyperparameter tuning for improving the efficacy of topic modeling in Arabic texts.

2. Related Work

Recent explorations into topic modeling have showcased its utility across various English corpora, affirming the effectiveness of methodologies such as NMF and LDA in distilling topics from extensive document collections [5,6]. The emergence of Bidirectional Encoder Representations from Transformers (BERT), a transformer model introduced by Devlin et al. [8], marks a significant advancement, broadening the scope of NLP tasks, including topic modeling.

A study by Bagheri, Entezarian, and Sharifi [9] explored the effectiveness of LDA, NMF, and BERTopic algorithms in identifying themes related to system thinking. Their research demonstrated the potential of these methods in extracting coherent and relevant topics from complex data sets. Additionally, it contributes significantly to the field by comparing traditional and advanced topic modeling techniques, underscoring practical implications for enhancing thematic analysis across various disciplines.

In 2022, Grootendorst introduced BERTopic [7], which uses a mix of document analysis and a special approach to identify topics better than previous methods. Through testing on with English news articles and comparing it to LDA, NMF, CTM, and Top2Vec, their study proved BERTopic's ability to create clear topics.

While these studies primarily focus on English, their insights are valuable for similar investigations in other languages, such as Arabic.

The introduction of AraBERT by Antoun, Baly, and Hajj [10] has significantly advanced Arabic NLP, demonstrating its superior performance in tasks such as sentiment analysis and named entity recognition. However, the application of AraBERT in topic modeling has yet to be fully explored, indicating a promising area for further research.

Ma, Al-Sabri, Zhang, Marah, and Al-Nabhan [11] examined the efficacy of topic modeling algorithms, specifically LDA and Dirichlet Multinomial Mixture (DMM), in processing Arabic texts. Their study focused on the impacts of various Arabic stemming processes on the efficiency of these algorithms. Through a comparative analysis of root-based, stem-based, and statistical approaches to stemming and evaluating the performance of models on documents processed with four different stemmers, their research shed light on the significant influence that stemming can have on the accuracy and coherence of the topics generated. Among the stemmers evaluated, the Farasa stemmer was highlighted for its effectiveness, which is ascribed to its advanced approach to navigating the morphological complexity of the Arabic language.

 doi:10.20944/preprints202411.1509.v1

Despite individual investigations into LDA, NMF, and BERTopic for Arabic NLP tasks, there is a noticeable lack of comparative analyses focused on these methods within the context of Arabic topic modeling. The following section highlights recent comparative studies in this area.

Topic Modeling Comparative Studies Using Arabic Data Sets

Despite the advancement of NLP studies on Arabic language data sets, the practical deployment of topic modeling for Arabic remains somewhat limited. This section delves into a selection of recent research, specifically concentrating on comparing topic modeling methods within the context of Arabic linguistics. In their research, Abdelrazek, Medhat, Gawish, and Hassan [12] conducted a comprehensive analysis comparing the efficacy of six topic modeling algorithms applied to a data set of Arabic newspaper articles. They established benchmarks using LDA and NMF models before comparing these with advanced neural topic models, such as the Contextualized Topic Model (CTM), Embedded Topic Model (ETM), AraBERT, and RoBERTa.

Their evaluation metrics included topic coherence, diversity, and computational efficiency. Remarkably, they found that the neural BERTopic framework, utilizing a RoBERTa-based sentence transformer, excelled in coherence, achieving a score 36% higher than BERTopic integrated with AraBERT, albeit with a 6% lower topic diversity compared to the CTM model and necessitating double the computational time. Their research also highlighted a pre-processing limitation, where the exclusive use of 1-gram tokens and noun retention posed challenges for neural models dependent on the corpus vocabulary size and document length. Addressing future research, the authors suggested broadening pre-processing approaches through including a wider array of token types and delving into n-gram models beyond $n = 1$. This would involve evaluating the impacts of different pre-processing strategies on neural model performance and customizing the steps to the specific needs of the model. They advocated for thoroughly evaluating these pre-processing modifications across various data sets to confirm the robustness and applicability of the findings, aiming to enhance the precision and relevance of topic extraction from Arabic texts and mitigate the initial study's pre-processing limitations.

The study by Abuzayed and Al-Khalifa [13] presented a thorough examination of the effectiveness of BERTopic, utilizing LDA and NMF performances as baselines for comparison. Employing a data set of 111,728 documents from three online newspapers, categorized into five distinct classes, their research aimed to measure topic coherence using Normalized Pointwise Mutual Information (NPMI) metrics. The study distinguished itself by leveraging various pre-trained language models, including monolingual Arabic BERT models (AraBERTV2.0, ARBERT, and QARiB) and a multilingual BERT model (XLM-R), demonstrating the superior accuracy and coherence of BERTopic in topic extraction compared to traditional methods. However, they fell short of providing comprehensive methodological details, such as the selection of n-grams and specific parameter settings for each algorithm. These omissions are pivotal for ensuring the study's reproducibility and for a full understanding of the basis on which the algorithms were evaluated. Furthermore, while the use of topic coherence as a primary evaluation metric is standard and valuable, the study's exclusive reliance on it without incorporating a wider array of metrics (e.g., perplexity and topic diversity) potentially narrowed the assessment of the algorithms' performance. Broadening the evaluation criteria to include multiple metrics would furnish a more nuanced and thorough analysis of each algorithm's capabilities.

Al-Khalifa et al. (2023), in their study titled "ChatGPT across Arabic Twitter: A Study of Topics, Sentiments, and Sarcasm" [14], explored the thematic landscape, sentiment orientation, and sarcasm detection among Arabic-speaking Twitter users interacting with ChatGPT. Analyzing a corpus of 34,760 tweets, the researchers identified predominant discussions centered around technological insights, ethical concerns, and regional applications of ChatGPT. In their approach to the research question concerning topic modeling, they examined the efficacy of various algorithms, including LDA, NMF, and BERTopic. They decided to proceed with BERTopic, due to its superior performance with Arabic documents. This method integrates transformer-based BERT embeddings with class-based TF-IDF

 doi:10.20944/preprints202411.1509.v1

to produce coherent and interpretable topics through document clustering based on semantic similarity. The model, utilizing an Arabic BERT-based embedding (aubmindlab/bert-base-arabertv02-twitter) and OpenAI's representation model (GPT-3.5 Turbo), was fine-tuned to delineate 50 topics via trial and error. It is crucial to emphasize that the data set was composed of Tweets, as this characteristic may yield distinct outcomes compared to handling longer texts.

Berrimi, Oussalah, Moussaoui, and Saidi (2023) [15] conducted a comparative study on Arabic text classification, evaluating 12 models, including 8 neural networks (LSTM, GRU, BiLSTM, BiGRU, CNN, C-BiLSTM, C-BiGRU, and Vanilla Transformer) and 4 pre-trained language models (AraBERT, ARBERT, MARBERT, and mBERT), and introducing a novel Transformer–CNN architecture. The models were assessed using accuracy, F1-score, and precision across multiple data sets. While the Transformer–CNN model showed superior performance, the pre-trained models—especially ARBERT—outperformed the neural network models, highlighting the effectiveness of pre-trained embeddings for Arabic text classification. The noted gaps for further research included a deeper investigation into the impact of pre-processing techniques on model performance.

3. Methodology

3.1. Data Set

This study utilizes the Single-label Arabic News Article Data set (SANAD), a comprehensive repository of 195,174 Arabic news articles sourced from three major outlets: AlKhaleej, AlArabiya, and Akhbarona [16]. The data set covers seven thematic categories: Culture, Finance, Medical, Politics, Religion, Sports, and Technology. The SANAD data set is chosen for its diversity and the detailed granularity it offers, making it a robust foundation for evaluating topic modeling techniques.

3.2. Pre-Processing Techniques

Pre-processing is crucial for refining Arabic text and enhancing the quality of topic modeling. Some of the techniques applied are detailed in the following.

- Text Normalization: Standardizing text by removing diacritics, numbers, punctuation, and special characters and unifying character variations [14].
- Tokenization: Splitting text into individual words or tokens, considering Arabic-specific linguistic rules [9].
- Stop Word Removal: Eliminating common Arabic words that do not contribute significant meaning [17].
- Stemming and Lemmatization: Reducing words to their root forms using two different stemmers—ISRI and Tashaphyne.
  - ISRI Stemmer: An Arabic stemmer based on the ISRI algorithm developed by the Information Science Research Institute. This stemmer removes prefixes, suffixes, and infixes from words and incorporates some degree of morphological analysis to accurately determine the roots of Arabic words. This approach generally enhances the accuracy of root identification [18].
  - Tashaphyne Stemmer: Provided by the Tashaphyne library, this is a light Arabic stemmer that merges light stemming and root-based stemming principles. It segments Arabic words into roots and patterns while also removing common prefixes and suffixes without performing a full morphological analysis. The Tashaphyne stemmer has achieved remarkable results, outperforming other competitive stemmers (e.g., Khoja, ISRI, Motaz/Light10, FARASA, and Assem stemmers) in extracting Roots and Stems [19].
- n-Gram Construction: Creating combinations of adjacent words (n-grams) to capture more context within the text [20].

 doi:10.20944/preprints202411.1509.v1

3.3. Modeling

The following three topic modeling techniques were evaluated.

- LDA: A probabilistic model that identifies topics based on word distributions within documents. It aims to infer these concealed topics by examining visible words [21]. Prior research has indicated the efficacy and utility of LDA methods in the field of topic modeling [22].
- NMF: A linear algebraic model that utilizes statistical techniques to identify thematic structures within a collection of texts. It employs a decompositional strategy based on matrix factorization, categorizing it within the realm of linear algebraic methods. It stands out as an unsupervised approach for simplifying the complexity of non-negative matrices. Additionally, when it comes to mining topics from brief texts, learning models based on NMF have proven to be a potent alternative to those predicated on LDA [23].
- BERTopic: A method for topic modeling that employs an unsupervised clustering approach. It leverages Bidirectional Encoder Representations from Transformers (BERT) to generate contextual embeddings of sentences. These embeddings encapsulate the semantic details of sentences, enabling the algorithm to identify topics through their contextual significance. The methodology of BERTopic encompasses a three-stage process. Initially, it transforms documents into embedding representations utilizing a pre-trained language model. Subsequently, it diminishes the dimensionality of these embeddings to facilitate more effective clustering. In the final stage, BERTopic derives topic representations from the clusters of documents through applying a unique class-based variant of TF-IDF, a technique detailed by Grootendorst in 2022 [7]. A key strength of BERTopic lies in its precision in forming clusters and its capability to propose names for topics based on these clusters. Unlike traditional topic modeling approaches, BERTopic does not require the number of topics to be pre-defined, offering a flexible and dynamic framework for topic discovery [24].

3.4. Evaluation Metrics

The models were evaluated using the following metrics.

- Topic Coherence: This metric measures the interpretability of the topics, with coherence scores calculated using Normalized Pointwise Mutual Information (NPMI) [25].
- Topic Diversity: This metric assesses the variety of topics through calculating the proportion of unique words among the top words across all topics [12].
- Perplexity: Used only for LDA, this metric assesses how well the model predicts a sample of the text [12].

4. Results

4.1. Training Experiments

This section presents the outcomes of the training experiments conducted with the three topic modeling techniques: LDA, NMF, and BERTopic. Each model was evaluated across four different data formats: Original articles without any pre-processing, articles after cleaning, articles after cleaning and stemming with the ISRI stemmer, and articles after cleaning and stemming with the Tashaphyne stemmer.

1. LDA Experiments: The results of the LDA experiments are summarized in Table 1.
    - Original Articles without Pre-processing: This baseline experiment resulted in a low coherence score (0.012) and moderate topic diversity (0.571). The model exhibited high perplexity (17,652.513), indicating difficulty in predicting the unseen data, as well as the longest training time (649.00 s).
    - Cleaning: After cleaning the text, the coherence score improved significantly to 0.040 and topic diversity increased to 0.729. The perplexity decreased to 14,655.042 and the training time was reduced to 380.82 s, reflecting the benefits of removing noise from the data.

 doi:10.20944/preprints202411.1509.v1

- Cleaning + Stemming (Tashaphyne): This configuration achieved the highest coherence score (0.051) among the LDA experiments, with a slight decrease in topic diversity (0.724). The perplexity dropped dramatically to 2836.581 and the training time decreased further to 289.69 s. This indicates that Tashaphyne stemming effectively enhanced the model's interpretability and prediction accuracy.
- Cleaning + Stemming (ISRI): While this setup resulted in a slightly lower coherence score (0.037) compared to Tashaphyne, it achieved the lowest perplexity (1424.293) among the LDA experiments, suggesting superior predictive performance. The training time was similar to the Tashaphyne setup, at 289.51 s.

Table 1. Summary of experimental results for the Latent Dirichlet Allocation (LDA) model.

Pre-processing	NPMI Coherence Score ↑	Topic Diversity	Perplexity
Original articles (without pre-processing)	0.012	0.571	17,652.513
Cleaning	0.040	0.729	14,655.042
Cleaning + stemming (Tashaphyne)	**0.051**	0.724	2836.581
Cleaning + stemming (ISRI)	0.037	0.643	1424.293

Among the LDA experiments, Experiment 3—with basic cleaning and stemming using Tashaphyne—achieved the highest coherence score (0.051) and good topic diversity (0.724). However, when considering predictive performance, Experiment 4—cleaning + stemming (ISRI)—stood out with the lowest perplexity. This suggests that, while stemming using Tashaphyne can enhance the interpretability of topics, incorporating ISRI stemming significantly improves the model's ability to predict and generalize to unseen data.

2. NMF Experiments: NMF was also tested under similar pre-processing configurations, with the results summarized in Table 2.

- Original Articles without Pre-processing: The NMF model initially achieved a coherence score of 0.032 and a lower topic diversity (0.557) than LDA. The training time was 2741.79 s, indicating that the model struggled with raw data.
- Cleaning: After cleaning, the coherence score improved to 0.038 and the topic diversity increased to 0.624. The training time significantly decreased to 1084.76 s, demonstrating that cleaning had a substantial positive impact.
- Cleaning + Stemming (Tashaphyne): This configuration resulted in the highest coherence score (0.071) and the best topic diversity (0.714) across all NMF experiments. The training time was further reduced to 528.50 s. The Tashaphyne stemmer clearly improved the model's ability to produce coherent and diverse topics.
- Cleaning + Stemming (ISRI): The ISRI stemming method also improved coherence (0.053) and diversity (0.681), although not as much as Tashaphyne. The training time was the lowest among all NMF experiments, at 396.90 s, highlighting ISRI's efficiency.

Table 2. Summary of experimental results for the Non-Negative Matrix Factorization (NMF) model.

Pre-processing	NPMI Coherence Score ↑	Topic Diversity
Original articles (without pre-processing)	0.032	0.557
Cleaning	0.038	0.624
Cleaning + stemming (Tashaphyne)	**0.071**	0.714
Cleaning + stemming (ISRI)	0.053	0.681

 doi:10.20944/preprints202411.1509.v1

Experiment 3, which involved cleaning and stemming using Tashaphyne, provided the highest coherence score (0.071) and topic diversity (0.714), making it the preferred choice for this technique. This indicates that Tashaphyne stemming significantly enhances the model's ability to produce coherent and diverse topics. Although Experiment 4 (with ISRI stemming) also improved the performance, it did not match the results achieved with Tashaphyne stemming. The Tashaphyne stemmer outperformed the ISRI stemmer primarily because ISRI's robust approach can occasionally lead to over-stemming. This over-stemming excessively reduces words, eliminating critical morphological details essential for differentiating between topics. Such aggressive reduction can create a diverse array of topics that are less meaningful or coherent, as more words are erroneously treated as identical. In contrast, Tashaphyne employs a lighter stemming technique, which likely avoids these issues. Through preserving more of the original word form, Tashaphyne ensures greater accuracy and relevance in the words retained, thereby enhancing the coherence of the topics generated.

3. BERTopic Experiments: Various embedding models were tested using different pre-processing techniques, and the results are summarized in Table 3.

    - UBC-NLP/MARBERT Embedding: Basic cleaning yielded the highest coherence score (0.110) and excellent topic diversity (0.871). Interestingly, stemming with ISRI slightly improved the topic diversity (0.886) but reduced coherence (0.046). Tashaphyne stemming resulted in a negative coherence score (-0.004), despite high topic diversity (0.857), suggesting that this stemmer may overly simplify the text for this embedding.
    - xlm-roberta-base Embedding: Cleaning significantly improved the coherence to 0.107 and maintained high topic diversity (0.843). Stemming with Tashaphyne and ISRI improved diversity but slightly reduced coherence, indicating that, while stemming is beneficial, it might not always enhance coherence with this embedding.
    - aubmindlab/bert-base-arabertv02 Embedding: Cleaning provided the best results, with a coherence score of 0.120 and topic diversity of 0.843. Stemming generally improved diversity, but had mixed effects on coherence.
    - CAMeL-Lab/bert-base-arabic-camelbert-da Embedding: Cleaning produced the highest coherence score (0.211) and the best topic diversity (0.886) across all BERTopic experiments. Stemming, while still effective, did not surpass the results achieved with basic cleaning.
    - asafaya/bert-base-arabic Embedding: Both cleaning and stemming methods performed well, with cleaning achieving a coherence score of 0.120 and topic diversity of 0.886. The original articles (without pre-processing) had slightly higher coherence (0.126) but lower diversity (0.843).
    - qarib/bert-base-qarib Embedding: With this embedding, cleaning provided the highest coherence (0.144) and excellent topic diversity (0.900). Stemming also performed well, maintaining high diversity and coherence close to that obtained after cleaning.

 doi:10.20944/preprints202411.1509.v1

Table 3. Summary of experimental results for the BERTopic model.

	Embedding Model	Pre-processing	NPMI Coherence Score ↑	Topic Diversity
1	UBC-NLP/MARBERT	Original articles (without pre-processing)	0.062	0.814
2		Cleaning	0.110	0.871
3		Cleaning + stemming (Tashaphyne)	−0.004	0.857
4		Cleaning + stemming (ISRI)	0.046	0.886
5	xlm-roberta-base	Original articles (without pre-processing)	−0.088	0.743
6		Cleaning	0.107	0.843
7		Cleaning + stemming (Tashaphyne)	0.104	0.800
8		Cleaning + stemming (ISRI)	0.083	0.829
9	aubmindlab/ bert-base-arabertv02	Original articles (without pre-processing)	0.024	0.757
10		Cleaning	0.120	0.843
11		Cleaning + stemming (Tashaphyne)	0.074	0.800
12		Cleaning + stemming (ISRI)	0.022	0.771
13	CAMeL-Lab/bert-base/ arabic/camelbert-da	Original articles (without pre-processing)	−0.040	0.857
14		Cleaning	**0.211**	0.886
15		Cleaning + stemming (Tashaphyne)	0.139	0.843
16		Cleaning + stemming (ISRI)	0.062	0.814

 doi:10.20944/preprints202411.1509.v1

Table 3. *Cont.*

	Embedding Model	Pre-processing	NPMI Coherence Score ↑	Topic Diversity
17	asafaya/bert-base-arabic	Original articles (without pre-processing)	0.126	0.843
18		Cleaning	0.120	0.886
19		Cleaning + stemming (Tashaphyne)	0.089	0.871
20		Cleaning + stemming (ISRI)	0.081	0.871
21	qarib/bert-base-qarib	Original articles (without pre-processing)	−0.026	0.786
22		Cleaning	0.144	0.900
23		Cleaning + stemming (Tashaphyne)	0.138	0.900
24		Cleaning + stemming (ISRI)	0.131	0.871

 doi:10.20944/preprints202411.1509.v1

4.2. Summary of Results

Based on an analysis of the 32 experiments, with each utilizing different topic modeling techniques and pre-processing steps, we can conclude that BERTopic emerged as the best-performing model. Among the various embeddings tested for BERTopic, the CAMeL-Lab/bert-base-arabic-camelbert-da embedding and the qarib/bert-base-qarib embedding stand out for achieving high coherence and topic diversity, specifically after applying cleaning to the data set, as can be observed from Figure 1. These results highlight the significance of pre-processing techniques in enhancing model performance. Further improvements could be achieved by exploring hyperparameter tuning.

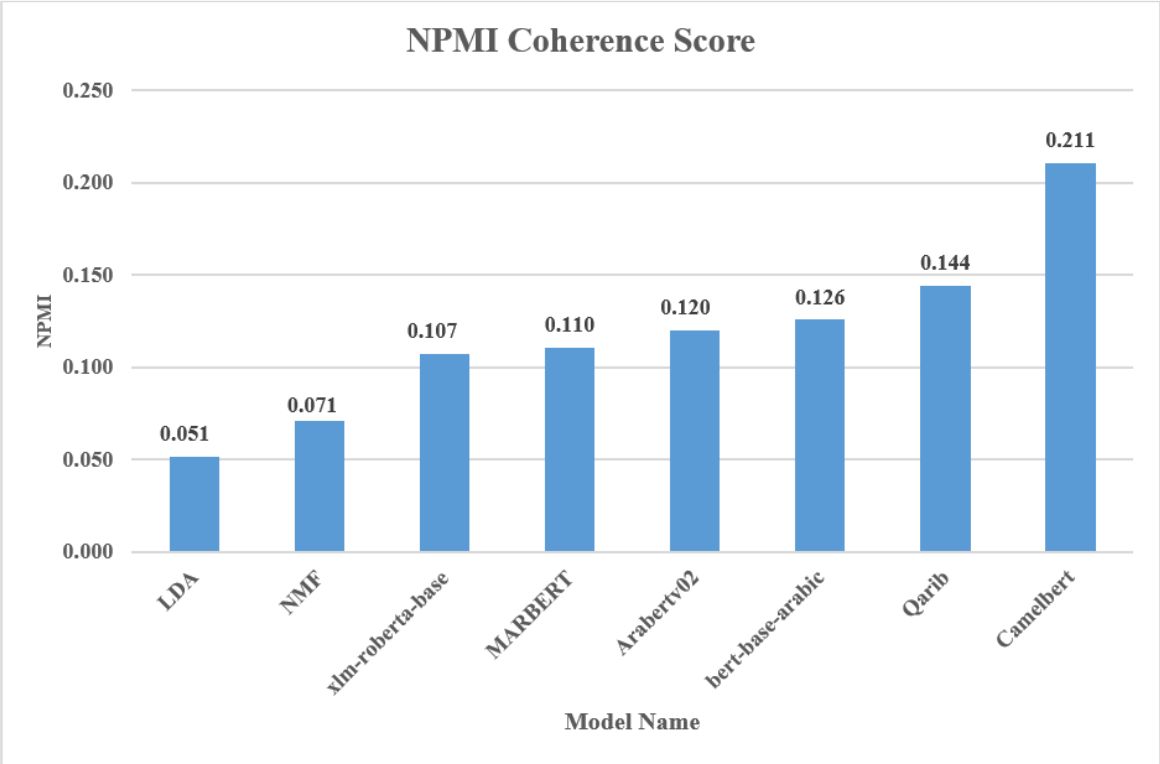

Figure 1. Normalized Pointwise Mutual Information (NPMI) coherence score.

4.3. Evaluation and Optimization

The best models identified from the previous experiments were tested using a held-out testing data set. To further optimize performance, we adjusted the `n_gram_range` parameter to include 2- and 3-grams. Figure 2 shows the results of testing before and after optimization. The results indicated that, while incorporating 3-grams slightly improved the NPMI score, it negatively impacted the overall performance. An analysis shows that the CAMeL model configured with 2-grams achieved the best balance, attaining high NPMI scores while maintaining efficient and acceptable performance. This suggests that 2-grams offer a more optimal configuration for achieving coherence and efficiency in the topic modeling task, when compared to 3-grams.

 doi:10.20944/preprints202411.1509.v1

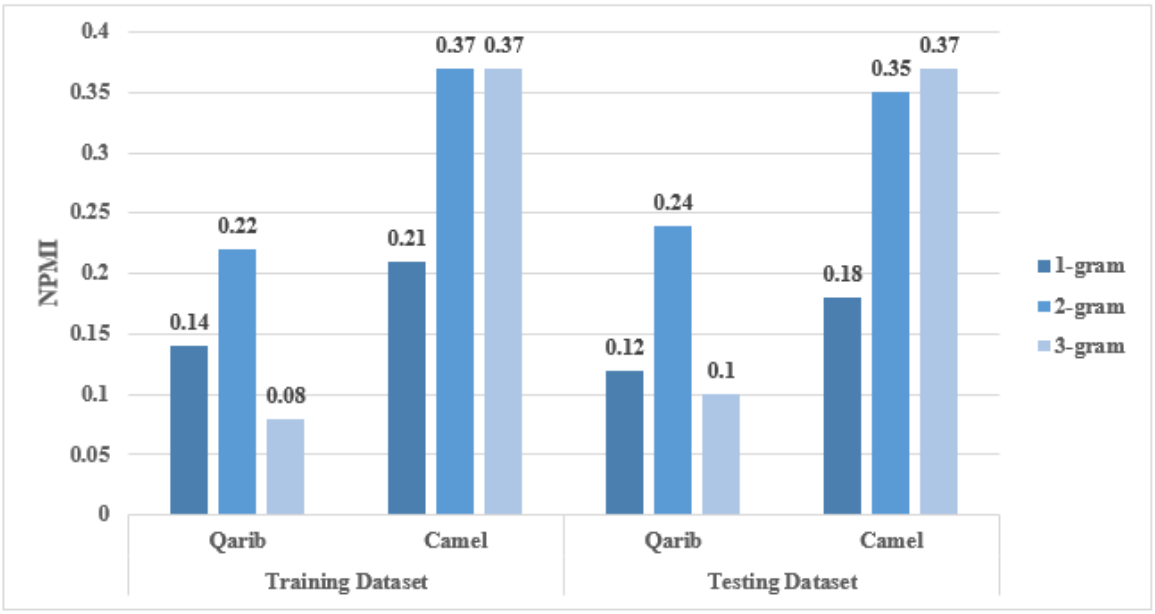

Figure 2. Evaluation and optimization results.

5. Conclusions

This study comprehensively explored the impacts of various pre-processing strategies on the effectiveness of BERTopic in modeling topics in Arabic texts, comparing it with the traditional methods LDA and NMF. The methodologies employed, the selection of the SANAD data set, and the detailed experimental setup provided a framework for evaluating the efficacy of these topic modeling techniques.

The comprehensive evaluation of the three models highlighted the significance of pre-processing techniques, particularly Tashaphyne stemming, in enhancing the performance of traditional topic modeling algorithms (i.e., LDA and NMF). For BERTopic, the combination of cleaning, employing 2-grams, and using the CAMeL-Lab/bert-base-arabic-camelbert-da embedding yielded the best NPMI score. This optimized approach achieved superior results, when compared to previous studies conducted by Abdelrazek, Medhat, Gawish, and Hassan [21], as well as Abuzayed and Al-Khalifa [13], both of which used similar Arabic news article data sets [26]. Specifically, this study achieved an NPMI score of 0.35, compared with their scores of 0.11 and 0.17, respectively, which translates to a 9% enhancement in topic coherence. In conclusion, this research demonstrated that BERTopic, with effective pre-processing and fine-tuning, is a powerful tool for topic modeling in Arabic texts. The insights gained from comparing BERTopic with traditional methods provide a valuable benchmark for future studies. Future research could explore the integration of large language models, such as Llama, for topic representation. This could enhance the richness and accuracy of topic models, leading to more sophisticated topic modeling applications.

However, this study has certain limitations. Firstly, Arabic is a highly diglossic language, with Modern Standard Arabic (MSA) used in formal settings and various regional dialects in informal contexts. The dataset used in this study focused solely on MSA, meaning that the findings may not generalize to Arabic dialects, which exhibit notable syntactic, lexical, and morphological differences. Secondly, the field lacks well-established benchmarks and evaluation metrics specifically tailored for Arabic topic modeling, making it challenging to evaluate model performance consistently and compare results across studies. Further research is needed to establish Arabic-specific benchmarks that accurately reflect topic coherence and interpretability. Lastly, topic modeling in Arabic remains a relatively under-researched field. While progress has been made, many methodologies and optimizations developed for other languages may not directly transfer to Arabic due to its unique linguistic

 doi:10.20944/preprints202411.1509.v1

characteristics. This limitation restricts the availability of best practices and comparative studies that could guide researchers in selecting the most effective approaches for Arabic topic modeling.

Author Contributions: Conceptualization, H.A. and N.A.; methodology, H.A.and N.A; software, H.A.; validation, H.A., N.A.; formal analysis, H.A and N.A.; investigation, H.A.and N.A.; resources, H.A. and N.A; data curation, H.A.; writing—original draft preparation, H.A.and N. A.; writing—review and editing, H.A. and N.A.; visualization, H.A. and N.A.; supervision, N.A.; project administration, H.A. and N.A.; funding acquisition, N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

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

References

1. Li, X.; Lei, L. A bibliometric analysis of topic modelling studies (2000–2017). *J. Inf. Sci.* **2021**, *47*, 161–175.
2. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1–22.
3. Blei, D.M.; Lafferty, J.D. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh , PA, USA, 25–29 June 2006; pp. 113–120.
4. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407.
5. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
6. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **2000**, *13, 535–541* .
7. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
8. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Bagheri, R.; Entezarian, N.; Sharifi, M.H. Topic Modeling on System Thinking Themes Using Latent Dirichlet Allocation, Non-Negative Matrix Factorization and BERTopic. *J. Syst. Think. Pract.* **2023**, *2*, 33–56 .
10. Antoun, W.; Baly, F.; Hajj, H. Arabert: Transformer-based model for arabic language understanding. *arXiv* **2020**, arXiv:2003.00104.
11. Ma, T.; Al-Sabri, R.; Zhang, L.; Marah, B.; Al-Nabhan, N. The impact of weighting schemes and stemming process on topic modeling of arabic long and short texts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2020**, *19*, 1–23.
12. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131.
13. Abuzayed, A.; Al-Khalifa, H. BERT for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Comput. Sci.* **2021**, *189*, 191–194.
14. Al-Khalifa, S.; Alhumaidhi, F.; Alotaibi, H.; Al-Khalifa, H.S. ChatGPT across Arabic Twitter: A Study of Topics, Sentiments, and Sarcasm. *Data* **2023**, *8*, 171.
15. Berrimi, M.; Oussalah, M.; Moussaoui, A.; Saidi, M. A Comparative Study of Effective Approaches for Arabic Text Classification. *SSRN Electronic Journal* **2023**.

    Available at SSRN: https://ssrn.com/abstract=4361591.
16. Einea, O.; Elnagar, A.; Al Debsi, R. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data Brief* **2019**, *25*, 104076.
17. El Kah, A.; Zeroual, I. The effects of pre-processing techniques on Arabic text classification. *Int. J.* **2021**, *10*, 1–12.
18. Taghva, K.; Elkhoury, R.; Coombs, J. Arabic stemming without a root dictionary. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II, Las Vegas, NV, USA, 4–6 April 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 152–157.

 doi:10.20944/preprints202411.1509.v1

19. Al-Khatib, R.M.; Zerrouki, T.; Abu Shquier, M.M.; Balla, A. Tashaphyne0. 4: A new arabic light stemmer based on rhyzome modeling approach. *Inf. Retr. J.* **2023**, *26*, 14.
20. Nithyashree, V. What Are N-Grams and How to Implement Them in Python? 2021. Available online: https://www.analyticsvidhya.com/blog/2021/11/what-are-n-grams-and-how-to-implement-them-in-python/ (accessed on 15 October 2024).
21. Abdelrazek, A.; Medhat, W.; Gawish, E.; Hassan, A. Topic Modeling on Arabic Language Dataset: Comparative Study. In Proceedings of the International Conference on Model and Data Engineering, Cairo, Egypt, 21–24 November 2022; Springer: Berlin, Germany, 2022; pp. 61–71.
22. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211.
23. Chen, Y.; Zhang, H.; Liu, R.; Ye, Z.; Lin, J. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl.-Based Syst.* **2019**, *163*, 1–13.
24. Egger, R.; Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498.
25. Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.; Blei, D. Reading tea leaves: How humans interpret topic models. *Adv. Neural Inf. Process. Syst.* **2009**, *22*.
26. Biniz, M. DataSet for Arabic Classification. *Mendeley Data*, *2*, 2018. https://doi.org/10.17632/v524p5dhpj.2.