Annals of Emerging Technologies in Computing (AETiC)
Vol. 9, No. 2, 2025

Research Article

Integration Named Entity Recognition and Latent Dirichlet Allocation to Enhance Topic Modeling

Hawraa Ali Taher*, Noralhuda N. Alabid and Bushra Mahdi Hasan

University of Kufa, Iraq
hawraaa.alshimirty@uokufa.edu.iq; noralhudan.hadi@uokufa.edu.iq; bushram.alhashimi@uokufa.edu.iq
*Correspondence: hawraaa.alshimirty@uokufa.edu.iq

Received: 8th July 2024; Accepted: 29th March 2025; Published: 1st April 2025

Abstract: Topic modeling from texts is one of the important topics in natural language processing (NLP), as it plays a fundamental role in summarizing texts, understanding their content, and facilitating access to the main ideas, especially in light of the vast quantity **of unstructured texts available today. Extracting titles is used in a variety of fields, such as news archiving, document classification, and content analysis in social media, making it an essential tool for improving information management and effective presentation. In this research, we focused on improving the methodology for extracting titles from texts by integrating two leading techniques: the topic assignment model using Latent Dirichlet Allocation (LDA) and the named entity recognition technique (NER). This combination aims to achieve a balance between identifying general topics of texts via LDA and extracting important information and key entities using NER, ensuring the generation of accurate and understandable titles that better reflect the actual content of the texts. The results of the study showed that the combined methodology achieved an accuracy of 71.97%, outperforming the performance of each technique separately, where the accuracy of NER alone was 29.78% and the accuracy of LDA alone was 67.80%. These results underscore the importance of integrating different techniques into NLP to improve headline extraction performance. This approach contributes to the development of more efficient text analysis methods, which enhances NLP applications in areas such as news analysis, content management, and document summarization, highlighting the importance of the topic in improving the handling of large texts and presenting them in a clearer and more appropriate way.**

Keywords: ***Bert Model; Cosine Similarity; Latent Dirichlet Allocation (LDA); Name Entity Recognition (NER); Text Analysis; Topic Modeling***

1. Introduction

Finding the subjects in brief texts like tweets and instant messages has grown in importance for a lot of content analysis software. The primary cause is that traditional topic models suffer from severe data sparsity in short documents because they implicitly capture document-level word co-occurrence patterns to reveal topics [1]. It has been demonstrated that methods like latent-semantic indexing and probabilistic topic models are generally helpful for dimension-reduction of sparse count data, as well as for automatically extracting the topical or semantic content of documents. It is possible to think of these kinds of models and algorithms as producing an abstraction from a document's words to a lower-dimensional latent variable representation that captures the general idea of the document in addition to the words that it contains [2].

As online communications progress, user comments a sizable collection of brief online texts are showing up more frequently. Its co-occurrence with typically longer normal documents is what distinguishes these brief texts. As an illustration, a news article or blog post may receive several user comments or reader reviews. Although it is rarely represented by traditional topic models, the co-occurring structure present in these text corpora is crucial for effective topic learning [3].

Hawraa Ali Taher , Noralhuda N. Alabid and Bushra Mahdi Hasan "Integration Named Entity Recognition and Latent Dirichlet Allocation to Enhance Topic Modeling", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 9, No. 2, 1st April 2025, pp. 20-30, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.02.002, Available: http://aetic.theiaer.org/archive/v9/v9n2/p2.html.

A topic model is an automatic tool for automatically organizing, comprehending, and summarizing large amounts of textual data. It is used in information retrieval to infer the hidden themes in a collection of documents. Additionally, topic models provide an interpretable representation of the documents that are utilized in various NLP tasks that are performed later on. There are several modeling approaches, ranging from the more modern neural models to probabilistic graphical models [4]. As science and technology advance at a rapid rate, the field of artificial intelligence is now characterized by a broad cross-disciplinary overlap, quick updates, and a renewed focus on international competition. Artificial intelligence is an interdisciplinary field with rich knowledge and strategic management significance. with regard to the topic recognition research [5].

Finding consistent and varied topics (sports, education, politics, etc.) within texts is the goal of topic modeling, a crucial text analysis technique that has been widely successful for many years in a variety of downstream data mining and natural language processing (NLP) tasks. The interpretable topics that topic models identify represent the decoupled elements that are necessary for the different text-processing tasks that come after. Topic modeling is commonly employed as an unsupervised technique in fundamental natural language processing tasks, like text classification and text clustering based on identified latent topics. Furthermore, neural topic modeling has proven beneficial for topic-driven text summarization as well as text generation. Traditional probabilistic topic models, which are mainly derived from Latent Dirichlet Allocation (LDA), include two probabilistic distributions of sampling words from topics and sampling topics from documents [6].

2.Related Work

Topic modelling is a fundamental technique in natural language processing (NLP) that aims to discover hidden patterns and topics within large collections of unstructured texts. Many previous studies have addressed the development and improvement of topic modeling models for various applications such as text classification, content analysis, and trend detection in text data. This section reviews the most prominent of these studies, focusing on methods to improve the accuracy of the models and their practical applications.

Rahimi *et al*. [7] proposed the Aligned Neural Topic Models (ANTM) algorithmic family of dynamic topic models, which combines cutting-edge data mining algorithms to offer a modular framework for finding emerging topics. By using advanced pre-trained Large Language Models (LLMs) to extract time-aware features from documents and sequentially clustering documents using an overlapping sliding window algorithm, ANTM preserves the temporal continuity of evolving topics. This overlapping sliding window algorithm aligns semantically similar document clusters across time periods and finds a distinct number of topics within each time frame. This method enables a more comprehensible portrayal of developing subjects by capturing emerging and fading trends over time. Three datasets used for evaluated ANTM against four other dynamic topic models.

Abdelrazek *et al.* [4] used four different topic models. The first part divides topic modeling approaches into four groups: neural, fuzzy, probabilistic, and algebraic. Examine the vast range of models that are available from each category, use a unified viewpoint to highlight the distinctions and similarities between models and model categories, look into the features and constraints of these models, and talk about appropriate use cases. Six criteria for appropriately evaluating topic models are shown in the second aspect, ranging from modeling quality to interpretability, stability, efficiency, and more. Because topic modeling is interpretable, it has been used in many different fields of study. We look at these programs as well as a few well-known software programs that offer model implementations. The fourth section examines benchmarks and datasets that are accessible.

Wang *et al.* [8] introduced a deep NMF (DNMF) topic modeling framework. In order to learn the latent hierarchical structures of documents, it first uses an unsupervised deep learning method. This approach is based on the idea that the topic word discovery problem can be improved if we can learn a good representation of documents, such as a deep model. After that, the deep model's output is used to constrain a topic-document distribution in order to find the discriminant topic words. This process not only increases efficacy but also lowers computational complexity when compared to traditional unsupervised NMF methods.

George and Sumathy [9] demonstrate a hybrid model of Bidirectional Encoder Representations from Transformers (BERT) and Latent Dirichlet Allocation (LDA) has been thoroughly examined in topic modeling with clustering based on dimensionality reduction, Due to the computational complexity of the clustering algorithms, which rises with the number of features, dimensionality reduction techniques based on PCA, t-SNE, and UMAP are also used. this study proposed a unified clustering-based framework (BERT and LDA) for mining a set of meaningful topics from the large text corpora.

Gao *et al.* [10] developed a new Topic-Aware BERT (TABERT) model, which extracts tweets' latent topics by first using a topic model. Second, topic data is combined with BERT's output using an adaptable framework. In the end, we use adversarial training to accomplish semi-supervised learning, and a significant amount of unlabelled data can be utilized to enhance the model's internal representations. Based on experimental results on the COVID-19 English tweet dataset, our model performs better than both traditional and cutting-edge baselines.

Koltcov *et al.* [11] developed topic modeling by testing four well-known and available to a wide range of user's topic models such as the embedded topic model (ETM), Gaussian Softmax distribution model (GSM), Wasserstein autoencoders with Dirichlet prior (W-LDA), and Wasserstein autoencoders with Gaussian Mixture prior (WTM-GMM). We demonstrate that W-LDA, WTM-GMM, and GSM possess poor stability that complicates their application in practice. ETM model with additionally trained embeddings demonstrates high coherence and rather good stability for large datasets, but the question of the number of topics remains unsolved for this model. We also propose a new topic model based on granulated sampling with word embeddings (GLDAW), demonstrating the highest stability and good coherence compared to other considered models. Moreover, the optimal number of topics in a dataset can be determined for this model.

Table 1. Comparison with the other published works

Author	**Algorithm**	**Data set Language**	**Contributions**
Hamed Rahimi *et al.* (2024) [7]	Neural Topic Models (ANTM) algorithmic	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques
Abdelrazek *et al.* (2023) [4]	neural, fuzzy, probabilistic, and algebraic models	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques
Wang *et al.* (2023) [8]	uses an unsupervised deep learning method	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques
George and Sumathy (2023) [9]	clustering-based framework (BERT and LDA)	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques
Gao *et al.* (2023) [10]	BERT (TABERT) model	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques
Koltcov *et al.* (2024) [11]	model (ETM), Gaussian Softmax distribution model (GSM), Wasserstein autoencoders with Dirichlet prior (W-LDA),	English	Clustering between (Named Entity Recognition and Latent Dirichlet Allocation) Techniques

3. Methodology and System Architecture

We investigate the topic-based hierarchical organization of large text databases to facilitate improved browsing, searching, and filtering. A lot of corpora are manually arranged into topic hierarchies, also known as taxonomies, including digital libraries, patent databases, and online directories. Like relational data indices, taxonomies improve search and access effectiveness. Nevertheless, the rate at which online textual data is growing exponentially makes it practically hard to manually maintain this kind of taxonomic organization for sizable, rapidly evolving corpora [12-14]. We describe an automatic, topics are automatically assigned to these new documents with a high degree of speed and accuracy. In this paper, use combining two technologies: topic assignment model using Latent Dirichlet Allocation (LDA), and named entity recognition (NER) technology for topic modeling.

3.1. Latent Dirichlet Allocation (LDA)

In order to determine topics within unstructured textual data, topic modeling is a machine learning technique that is widely used in Natural Language Processing (NLP) applications. One of the most popular topic modeling methods is Latent Dirichlet Allocation (LDA), which can automatically identify topics from

a sizable collection of text documents. Originally created by Blei *et al.* (2003), latent Dirichlet allocation (LDA) is used to expose the corpus's hidden semantic structure [6]. Nowadays, latent topic themes in a set of documents can be found using the three-layer Bayesian model known as Latent Dictionary Analysis (LDA). The primary concept is that every document displays a variety of latent topics, each of which is identified by a probability distribution across the words (individual markers for the set of documents). One well-liked method for locating latent topics in a corpus of text documents is Latent Dirichlet Allocation (LDA) [15]. Using the Count Vectorizer, we first apply text vectorization to transform the text data into a numerical format appropriate for modeling. Next, we eliminated common stop words in English and specified parameters like the maximum and minimum document frequency and the maximum number of features (words) to take into account. Next, we used the Latent Dirichlet Allocation model as an example to apply LDA to five topics. The distribution of each article among these five topics is shown in the topic matrix that results. Next, we designate a dominant topic for every article by selecting the topic with the greatest likelihood [16]. An algorithm 1, that explains how LDA to determine the appropriate topic. This algorithm is summarized in the following steps:

- Determine the total number of words in the article
- Determine the distribution ratios of topics in the article
- Clustering topics in an unsupervised depending on the presence of similar words in it, using an algorithm LDA.

Algorithm 1. Latent Dirichlet Allocation to Discover 5 Topics

1. Initialize Parameters:
1.1 Set the number of topics K=5
1.2 Choose Dirichlet hyper parameters: α (topic distribution per document) and β (word distribution per topic).
2. Assign Initial Topics:
Randomly assign a topic Z_{dn} to each word W_{dn} in each document d
3. Iterate Gibbs Sampling Steps:
3.1 Remove the current assignment

$$N_{dk} = N_{dk} - 1 \quad (1)$$

$$N_{kv} = N_{kv} - 1 \quad (2)$$

3.2 Calculate conditional probability:

$$P(Z_{dn} = K|Z_{-dn}, W) \propto \frac{N_{dk-dn}+\alpha_k}{\sum_{k'}(N_{dk'-dn}+\alpha_{k'})} \cdot \frac{N_{kv-dn}+\beta_v}{\sum_{v'}(N_{kv'-dn}+\beta_{v'})} \quad (3)$$

3.3 Assign a new topic.
3.4 choose new topic Zdn for the word based on the calculated probabilities.
3.5 Update counts:

$$N_{dk} = N_{dk} + 1 \quad (4)$$

$$N_{kv} = N_{kv} + 1 \quad (5)$$

4. Repeat:
4.1 Repeat the Gibbs Sampling step for many iterations until topic assignment stabilize.
5. Estimate Final Parameters:
5.1 Document-Topic Distribution (θ):

$$\theta_{dk} = \frac{N_{dk} + \alpha_k}{\sum_{k'}(N_{dk'} + \alpha_{k'})} \quad (6)$$

5.2 Topic-Word Distribution (ϕ):

$$\phi_{kv} = \frac{N_{kv} + \beta_v}{\sum_{v'}(N_{kv'} + \beta_{v'})} \quad (7)$$

6. Extract Topics:
6.1 Use θ and ϕ to identify the main topics and the most relevant words for each topic.

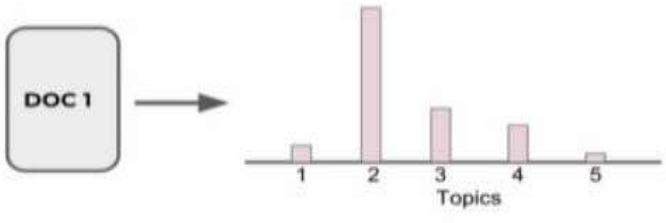

Figure 1. Generalized structure of LDA for Topic Model in Document

3.2. Name Entity Recognition (NER) Model

NER, or name entity recognition, is a crucial natural language processing subtask. In the past ten years, numerous NER systems have been developed. They might focus on various domains, use various techniques, work with various languages, identify various entity kinds, and accommodate various input and output formats. It is challenging for a user to choose the appropriate NER tools for a given task under these circumstances. NER is a natural language processing technique that allows you to extract and identify particular entities from the text, including names, dates, locations, organizations, and more. Each article is analyzed by the extract named entities function, which then classifies the entities according to their labels (e.g., "ORG" for organizations and "LOC" for locations). NER, or name entity recognition, is a crucial natural language processing subtask. In the past ten years, numerous NER systems have been developed. They might focus on various domains, use various techniques, work with various languages, identify various entity kinds, and accommodate various input and output formats. It is challenging for a user to choose the appropriate NER tools under these circumstances [17-18]. In our work, this technique will be applied to determine a title for a text or document. Figure 2, shown generalized structure of NER for topic model. And An algorithm 2, that explains how NER to determine the appropriate topic.

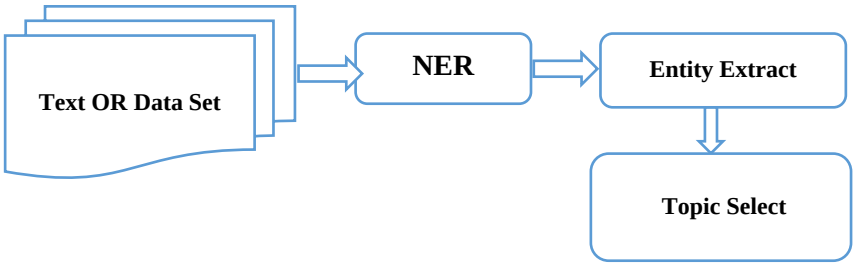

Figure 2. Generalized structure of NER for Topic Model

Algorithm 2. Named Entity Recognition for Topic Modeling

1. Load the NLP model:
   M=spacy.load('xx_ent_wiki_sm')
2. process the text:
   D(T)=M(T)
3. Extract Name Entity:
   E= {}
4. Count Occurrences and Determine the Top Title:
   C= {}
   for e_type,e_list in E:
   for e in e_list:
   C[e]+=1
   Top title= arg max C(e)
5. print (Top title)

3.3. Integrating NER and LDA for Topic Model

Natural language processing (NLP) algorithms are essential tools for extracting and organizing information from unstructured texts. The combination of named entity recognition (NER) and topic modeling using latent Dirichlet allocation (LDA) techniques is a notable innovation aimed at improving text analysis and topic discovery [19]. Combining Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) can enhance topic classification in texts by focusing on important entities within the text. This algorithm uses NER to extract named entities from the texts and then uses LDA to identify topics based on these entities. An algorithm 3, that explains how integrating NER and LDA to determine the appropriate topic.

Algorithm 3. Integrating NER and LDA for Topic Modeling

1. Load Data
2. Text Processing
3. NER Mode for extract Named Entities from text

$$NER(t_i) = \{e_1, e_2, .., e_k\} \quad (8)$$

   Where t_i is input text and $e_1, e_2, .., e_k$ are extract Named Entities
4. Create Document –Term Matrix Using Named Entities and processed text:

$$X = Count\ Vectorizer(\{e_1, e_2, .., e_k\}) \quad (9)$$

5. Apply LDA Model on Document-Term Matrix

$$\phi, \theta = LDA(X, K) \tag{10}$$

Where X is the Document-Term Matrix., K is the number of topics, θ is the document-topic distribution, ϕ is the topic-word distribution.

6. Extract and Distribute Topics:
Each topic k can be represented by the words with the highest probabilities:

$$Topic_k = argsort(\phi_k)_{top\ n\ words} \tag{11}$$

7. Distribute Topics for Each Document:
Each document d can be represented by the topic distribution:

$$Documnet_d = \theta_d \tag{12}$$

4. The Evaluation

To evaluate the accuracy of the resulting titles using two models: Named Entity Extraction (NER) for extracting important titles, and Latent Dirichlet Allocation (LDA) for uncovering hidden topics in texts, we calculate the similarity between the embeddings [20]. These embeddings are extracted using BERT, which captures the contextual representation of texts, where each text is numerically represented to reflect its meaning within a specific context, each text is converted into numerical representations (Embeddings) using the BERT model, which takes into account the linguistic context of the words and produces vectors that represent the texts comprehensively [21]. The similarity between texts is then computed using the Cosine Similarity metric. The similarity between text, the similarity of two documents can be related to the relationship between the corresponding vectors of the documents. Perfect similarity defines the similarity according to the angle between the vectors where the most similar documents have small angles between their vectors. Perfect similarity is one of the most important criteria commonly used for text documents. Suppose that we have two document A and B and the two corresponding vectors $\vec{A}$ and$\vec{B}$ then the cosine distance between the two vectors $\vec{A}$ and $\vec{B}$. Which measures the directional similarity between the BERT-generated embeddings, aiding in determining how closely the texts are related or match the title [22-24]. An algorithm 4, that explains how calculating similarity score between texts using cosine similarity and BERT.

Algorithm 4. Calculating Similarity Score Between Texts Using Cosine Similarity and BERT

1. Tokenize Text:
Convert each input text into tokens using a tokenizer:

$$Tokens = Tokenize(Text) \tag{13}$$

2. Generate Embedding:
Pass the tokenized texts into a pre-trained BERT model to extract embeddings

$$Embeddings = BERT(Tokens) \tag{14}$$

3. Compute Sentence Embeddings:
Calculate the average of the embeddings for all tokens in each text to represent the entire text as a single vector.

$$Sentence\_Embedding = \frac{1}{N}\sum_{i=1}^{N} Token\_Embedding_i \tag{15}$$

Where N is the number of tokens.

4. Calculate Cosine Similarity:
Use the cosine similarity formula to compare the two sentence embeddings

$$Cosine\ Similarity = \frac{Embedding_1 . Embedding_2}{||Embedding_1|| \times ||Embedding_2||} \tag{16}$$

5. Return the Similarity Score:
Output the calculated cosine similarity score, which ranges between -1 and 1.

5. Experimental Results and Performance Analysis

The following the results of the proposed system for topic modeling for three of the datasets using LDA and NER, Figure (3), Figure (4), Figure (5) show important words in document using word cloud.

Figure 3. Word Cloud for Dataset1

Figure 4. Word Cloud for Dataset2

Figure 5. Word Cloud for Dataset3

Table (2), show results for name entity recognition for dataset1, dataset2, dataset3:

Table 2. Results of NER for Datatset1, Dataset2, Dataset3

No	Entity	Count for Dataset1	Count for Dataset2	Count for Dataset3
0	Data	8	2161	11781
1	Cardinal	9	1865	11861
2	ORG	20	1977	11153
3	Ordinal	1	1398	8471
4	LOC	1	444	11846

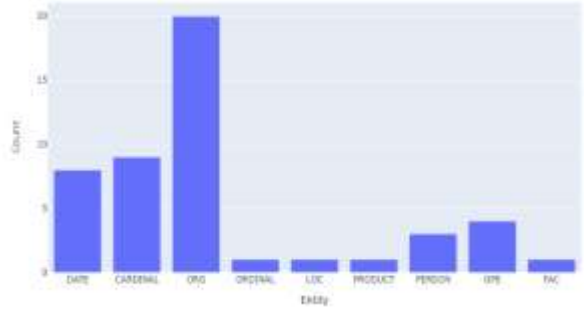

Figure 6. Results of NER for Dataset1

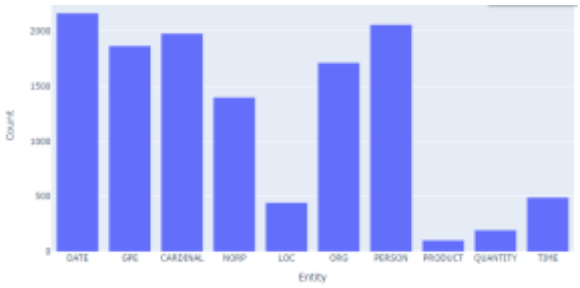

Figure 7. Results of NER for Dataset2

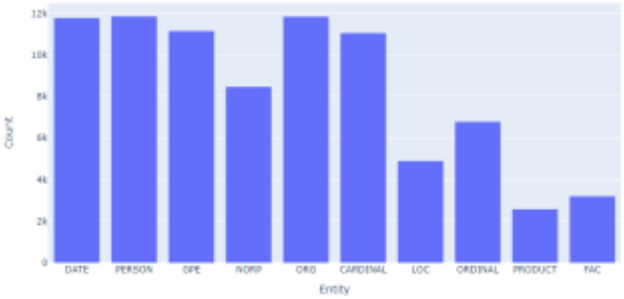

Figure 8. Results of NER for Dataset3

Figure (9), Figure (10) and Figure (11) show a bar chart that illustrates the distribution of recurrent themes in datasets 1, 2, and 3 to give readers a summary of the main themes or topics covered in the article.

Table 3. Results of LDA for Datatset1, Dataet2, Dataset3

No	Index	Dominant Topic for Dataset1	Dominant Topic for Dataset2	Dominant Topic for Dataset3
0	Topic 1	18	512	4758
1	Topic 0	6	502	2328
2	Topic 1	4	472	1795
3	Topic 3	3	377	1673
4	Topic 4	3	362	1438

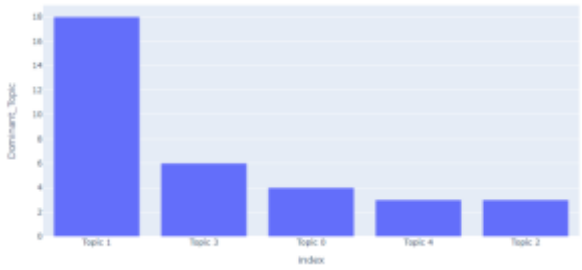

Figure 9. Results of LDA for Datatset1

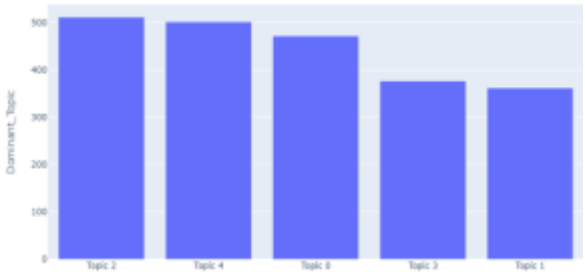

Figure 10. Results of LDA for Datatset2

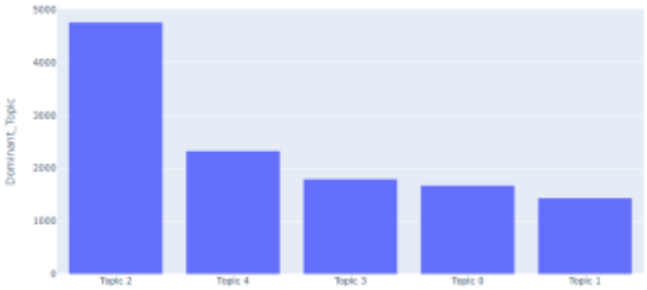

Figure 11. Results of LDA for Datatset3

As following Table4, are the results of specifying a title for the text based on NER for each text in dataset1 for 10 texts:

Table 4. Results of NER for topic Model for Datatset1

Number of texts	The Title
Text 0	Today
Text 1	No Title Found
Text 2	No Title Found
Text 3	only two
Text 4	The Multinomial Naive Bayes
Text 5	No Title Found
Text 6	NLP
Text 7	Instagram
Text 8	Pfizer
Text 9	NetFlix
Text 10	Computer Vision

The result "**No Title Found**" occurs when using NER due to the absence of relevant entities in the text, limitations of the algorithm in recognizing entities or context, poor quality of the input texts, or the lack of integration with other algorithms like LDA to provide deeper context and comprehensive analysis. Table 5, showing results of Integrating NER and LDA for Topic Model Datatset1

Table 5. Results of Integrating NER and LDA for Topic Model Datatset1

No	Index	Dominant Topic (LDA)	Dominant Topic (LDA and NER)
0	Topic 0	Applications of Deep Learning	machine learning algorithm clustering learn python classification news insurance cluster
1	Topic 1	Introduction to Recommendation Systems	datum deep learn learning network neural data value need use
2	Topic 2	Use Cases of Different Machine Learning Algorithms	computer good learn book artificial application intelligence look face example
3	Topic 3	Naive Bayes Algorithm in Machine Learning	python api today want learn application programming good tutorial datum
4	Topic 4	Swap Items of a Python List	learning machine bayes naive language stock algorithm price task detection

We notice from the results in Table (5), it involves processing the texts and merging them with the named entities, and then extracting the main topics based on the final matrix created from the processed texts and the named entities. The topic is determined based on the appearance of the most frequently repeated words. Table 6, showing of similarity score results for LDA and results for NER. And Table 7 showing accuracy results for LDA, NER and integrating NER and LDA.

The results show in the above Table (6) and Table (7), the combination of NER and LDA methods to improve the accuracy of the resulting addresses showed a clear superiority, as their combination achieved an accuracy of 71.97%, which is higher than using each method separately. The low performance of NER

alone, with an accuracy of 29.78%, reflects the method's limitation in providing comprehensive addresses due to its focus on named entities only. In contrast, the LDA method achieved a better accuracy of 67.80%, thanks to its ability to extract hidden topics, but it remains less accurate compared to the combination. This indicates that the combination of NER and LDA provides more accurate and comprehensive addresses than either of them alone.

Table 6. Similarity Score Results for LDA and Results for NER

Article number	Topic Distribution (LDA)	Similarity Score (LDA)	Similarity Score (NER)
Article1	[0.0027017862313356024, 0.9891964057814929, 0.002713724367259702, 0.0027033886279202667, 0.002684694991991455]	0.70	0.25
Article2	[0.0037831556569309236, 0.9849485398574604, 0.003754421492798723, 0.0038069772047022426, 0.0037069057881077075]	0.55	0.00
Article3	[0.0029106992630936342, 0.002927608680553026, 0.988402769464761, 0.0029007349353937805, 0.0028581876561986123]	0.75	0.00
Article4	[0.0031250629325322793, 0.98753249452777, 0.0031311072564852477, 0.0031248548932963503, 0.0030864803899161236]	0.67	0.31
Article5	[0.99143317377994, 0.002152696210425865, 0.002146462966025 4866, 0.0021615785002459783, 0.0021060885433626443]	0.64	0.69
Article6	[0.0029106992630936342, 0.002927608680553026, 0.988402769464761, 0.0029007349353937805, 0.0028581876561986123	0.75	0.00
Article7	[0.98983736202948, 0.002561064247826961, 0.0025459469496899386, 0.0025375599967564287, 0.0025180667762467386]	0.72	0.25
Article8	[0.0026954074862611185, 0.002710620552858503, 0.989225081272018, 0.0026991782107457177, 0.0026697124781165384]	0.70	0.41
Article9	[0.002504790301266291, 0.9900126123343392, 0.002511929802511304, 0.002500003628728046, 0.0024706639331552207]	0.70	0.41
Article10	[0.0023321446334528375, 0.9906846736717929, 0.0023362926338311957, 0.0023464618906936328, 0.0023004271702295256]	0.73	0.28

Table 7. Accuracy Results

Algorithm	**Accuracy**
System (NER) Accuracy	29.78%
System (LDA) Accuracy	67.80%
System (NER+LDA) Accuracy	71.97%

6. Conclusion

Information processing and filtering from a large number of documents many of which exist in digital form has become increasingly difficult due to the widespread use of information technology, the advancement of the World Wide Web, and its applications. These documents include academic reports, technical patents, social media posts, and online newspapers. The sheer size and rapid content updates of these documents make it challenging for users to sift through collections in search of relevant information. To address this, various techniques have been developed, such as topic modeling, multi-document summarization, segment-level extraction of pertinent passages, document-level retrieval, and semantic-level extraction of entities and relations. In this work, we focused on addressing the growing interest in finding topic-centric content by integrating Latent Dirichlet Allocation (LDA) for topic modeling with Named Entity Recognition (NER) for entity extraction. The combined approach was tested using three datasets to demonstrate its potential in generating meaningful titles and improving topic classification.

The integration of LDA and NER presents a novel contribution to the field of natural language processing (NLP), enabling the identification of topic-specific content while providing relevant, context-aware titles. This research highlights the benefits of combining these techniques, particularly in applications where extracting meaningful information from vast document collections is critical. Future enhancements to the proposed model could involve adopting more advanced NER models, expanding the approach to multilingual datasets, and incorporating deep learning techniques for improved accuracy and adaptability. Additionally, the versatility of the model makes it suitable for various domains, including social media analysis, market trend prediction, and academic research. By bridging topic modeling and entity recognition, this work contributes to the development of intelligent and context-aware NLP systems, paving the way for further innovations in processing and extracting meaningful insights from unstructured text data.

CRediT Author Contribution Statement

Hawraa Ali Taher: Writing-Original draft preparation, Writing –review & editing, Conceptualization, Supervision, Project administration, Software; Noralhuda N. Alabid: Methodology; Bushra Mahdi Hasan: Validation.

References

[1] Xiaobao Wu, Thong Nguyen and Anh Tuan Luu, "A survey on neural topic models: methods, applications, and challenges", *Artificial Intelligence Review*, Print ISSN: 0269-2821, Online ISSN: 1573-7462, Vol. 57, 25 January 2024, Published by Springer Nature, DOI: 10.1007/s10462-023-10661-7, Available: https://link.springer.com/article/10.1007/s10462-023-10661-7.

[2] Mekhail Mustak, Joni Salminen, Loïc Plé and Jochen Wirtz, "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda", *Journal of Business Research*, Print ISSN: 0148-2963, Online ISSN: 1873-7978, Vol. 124, January 2021, pp. 389-404, Published by Elsevier, DOI: 10.1016/j.jbusres.2020.10.044, Available: https://www.sciencedirect.com/science/article/abs/pii/S0148296320307165.

[3] Yang Yang, Feifei Wang, Junni Zhang, Jin Xu and Philip S. Yu, "A topic model for co-occurring normal documents and short texts", *World Wide Web*, Print ISSN: 1386-145X, Online ISSN: 1573-1413, Vol. 21, 23 June 2017, pp. 487–513, Published by Springer, DOI: 10.1007/s11280-017-0467-8, Available: https://link.springer.com/article/10.1007/s11280-017-0467-8.

[4] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat and Ahmed Hassan, "Topic modeling algorithms and applications: A survey", *Information Systems*, Print ISSN: 0306-4379, Online ISSN: 1873-6076, Vol. 112, February 2023, Published by Elsevier, DOI: 10.1016/j.is.2022.102131, P. 102131, Available: https://www.sciencedirect.com/science/article/abs/pii/S0306437922001090.

[5] Yunmei Liu and Min Chen, "The knowledge structure and development trend in artificial intelligence based on latent feature topic model", *IEEE Transactions on Engineering Management*, Print ISSN: 0018-9391, Online ISSN: 1558-0040, Vol. 71, 16 January 2023, pp. 12593-12604, Published by IEEE, DOI: 10.1109/TEM.2022.3232178, Available: https://ieeexplore.ieee.org/abstract/document/10017444.

[6] Xixi Zhou, Jiajun Bu, Sheng Zhou, Zhi Yu, Ji Zhao *et al.*, "Improving topic disentanglement via contrastive learning", *Information Processing & Management*, Print ISSN: 0306-4573, Online ISSN: 1873-5371, Vol. 60, No. 2, March 2023, Published by Elsevier, DOI: 10.1016/j.ipm.2022.103164, Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457322002655.

[7] Hamed Rahimi, Hubert Naacke, Camelia Constantin and Bernd Amann, "ANTM: Aligned Neural Topic Models for Exploring Evolving Topics", In *Lecture Note of Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI*, Vol. 14790**,** Online ISBN: 978-3-662-69603-3, Print ISBN: 978-3-662-69602-6, Series Print ISSN: 0302-9743, Series Online ISSN: 1611-3349, DOI: https://doi.org/10.1007/978-3-662-69603-3_3, 21 July 2024, pp. 76-97, Published by Springer Berlin Heidelberg, Available: https://link.springer.com/chapter/10.1007/978-3-662-69603-3_3.

[8] Jianyu Wang and Xiao-Lei Zhang, "Deep NMF topic modeling", *Neurocomputing*, Print ISSN: 0925-2312, Online ISSN: 1872-8286, Vol. 515, 1 January 2023, pp. 157-173, Published by Elsevier, DOI: 10.1016/j.neucom.2022.10.002, Available:https://www.sciencedirect.com/science/article/abs/pii/S0925231222012632.

[9] Lijimol George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling", *International Journal of Information Technology*, Print ISSN: 2511-2104, Online ISSN: 2511-2112, Vol. 15, 6 May 2023, pp. 2187–2195, Published by Springer Nature, DOI: 10.1007/s41870-023-01268-w, Available: https://link.springer.com/article/10.1007/s41870-023-01268-w.

[10] Wang Gao, Lin Li, Xiaohui Tao, Jing Zhou and Jun Tao, "Identifying informative tweets during a pandemic via a topic-aware neural language model", *World Wide Web*, Print ISSN: 1386-145X, Online ISSN: 1573-1413, Vol. 26, No. 1, 16 March 2022, pp. 55–70, Published by Springer Nature, DOI: 10.1007/s11280-022-01034-1, Available: https://link.springer.com/article/10.1007/s11280-022-01034-1.

[11] Sergei Koltcov, Anton Surkov, Vladimir Filippov and Vera Ignatenko, "Topic models with elements of neural networks: investigation of stability, coherence, and determining the optimal number of topics", *PeerJ Computer Science*, Print ISSN: 2376-5992, Online ISSN: 2376-600X, Vol. 10, 3 January 2024, Published by PeerJ Inc., DOI: 10.7717/peerj-cs.1758, Available: https://peerj.com/articles/cs-1758/.

[12] Uttam Chauhan and Apurva Shah, "Topic modeling using latent Dirichlet allocation: A survey", *ACM Computing Surveys (CSUR)*, Print ISSN: 0360-0300, Online ISSN: 1557-7341, Vol. 54, No. 7, 17 September 2021, pp.1-35, Published by Association for Computing Machinery, DOI: 10.1145/3462478, Available: https://dl.acm.org/doi/abs/10.1145/3462478.

[13] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Online ISSN: 1573-7721, Vol. 78, 28

November 2018, pp. 15169–15211, Published by Springer Nature, DOI: 10.1007/s11042-018-6894-4, Available: https://link.springer.com/article/10.1007/s11042-018-6894-4.

[14] Saad Ahmed Sazan, Mahdi H. Miraz and A B M Muntasir Rahman, "Enhancing Depressive Post Detection in Bangla: A Comparative Study of TF-IDF, BERT and FastText Embeddingss", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 8, No. 3, 1st July 2024, pp. 34-50, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2024.03.003, Available: http://aetic.theiaer.org/archive/v8/v8n3/p3.html.

[15] Jian Ma, Lei Wang, Yuan-Rong Zhang, Wei Yuan and Wei Guo, "An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local", *Expert Systems with Applications*, Print ISSN: 0957-4174, Online ISSN: 1873-6793, Vol. 212, February 2023, Published by Elsevier, DOI: 10.1016/j.eswa.2022.118695, Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417422017250.

[16] Amreen Batool and Yung-Cheol Byun, "Enhanced Sentiment Analysis and Topic Modeling During the Pandemic Using Automated Latent Dirichlet Allocation", *IEEE Access*, ISSN: 2169-3536, Vol. 12, 17 June 2024, pp. 81206-81220, Published by IEEE, DOI: 10.1109/ACCESS.2024.3411717, Available: https://ieeexplore.ieee.org/abstract/document/10552275.

[17] Zhentao Hu, Wei Hou and Xianxing Liu , "Deep learning for named entity recognition: a survey", *Neural Computing and Applications*, ISBN 978-1-945626-16-6, Vol. 36, No. 16, 28 March 2024, pp. 8995-9022, Published by Springer, DOI: https://doi.org/10.1007/s00521-024-09646-6, Available: https://link.springer.com/article/10.1007/s00521-024-09646-6.

[18] Behrang Mohit, "Named entity recognition", *In Natural Language Processing of Semitic Languages*, Print ISBN: 978-3-642-45357-1, Online ISBN: 978-3-642-45358-8, 25 March 2014, pp. 221-245, Published by Springer, DOI: 10.1007/978-3-642-45358-8_7, Available: https://link.springer.com/chapter/10.1007/978-3-642-45358-8_7.

[19] Diellza Nagavci Mati, Mentor Hamiti, Arsim Susuri, Besnik Selimi and Jaumin Ajdari, "Building Dictionaries for Low Resource Languages: Challenges of Unsupervised Learning", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 5, No. 3, 1st July 2021, pp. 52-58, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.03.005, Available: http://aetic.theiaer.org/archive/v5/v5n3/p5.html.

[20] Omar Galal, Ahmed H. Abdel-Gawad and Mona Farouk, "Rethinking of BERT sentence embedding for text classification", Neural *Computing and Applications*, Print ISSN: 0941-0643, Online ISSN: 1433-3058, Vol. 36, 12 August 2024, pp. 20245-20258, Published by Springer, DOI: 10.1007/s00521-024-10212-3, Available: https://link.springer.com/article/10.1007/s00521-024-10212-3.

[21] Noralhuda Alabid and Hawraa Ali Taher, "Enhancing Arabic fake news detection with a hybrid MLP-SVM approach and Doc2Vec embeddings ", *International Journal of Advanced Technology and Engineering Exploration,* Print ISSN : 2394-5443, Online ISSN: 2394-7454, Vol. 11, No. 121, December 2024, pp. 1732 -1746, Published by Accent Social and Welfare Society, DOI: 10.19101/IJATEE.2024.111100949, Available: https://www.accentsjournals.org/paperInfo.php?journalPaperId=1738&countPaper=114.

[22] Murat Kirişci, "New cosine similarity and distance measures for Fermatean fuzzy sets and TOPSIS approach", *Knowledge and Information Systems* , Print ISSN: 0219-1377, Online ISSN: 0219-3116, Vol. 65, No. 2, 4 November 2022, pp. 855–868, Published by Springer, DOI: 10.1007/s10115-022-01776-4, Available: https://link.springer.com/article/10.1007/s10115-022-01776-4.

[23] Tri Puspa Rinjeni, Ade Indriawan and Nur Aini Rakhmawati, "Matching scientific article titles using cosine similarity and jaccard similarity algorithm", *Procedia Computer Science*, Online ISSN: 1877-0509, Vol. 234, 29 April 2024, pp. 553-560, Published by Elsevier, DOI: 10.1016/j.procs.2024.03.039, Available: https://www.sciencedirect.com/science/article/pii/S1877050924003971.

[24] Takuro Hada, Yuichi Sei, Yasuyuki Tahara and Akihiko Ohsuga, "Codeword Detection, Focusing on Differences in Similar Words Between Two Corpora of Microblogs", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 5, No. 2, 1st April 2021, pp. 90-102, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.02.008, Available: http://aetic.theiaer.org/archive/v5/v5n2/p8.html.