Application of Topic Modeling Techniques in Arabic Content: A Systematic Review

Maram Alhmiyani and Huda Alhazmi,

Umm Al-Qura University, College of Computer and Information System, Makkah, Saudi Arabia

Abstract
With the rapid increase of user generated data on digital platforms, the task of categorizing and classifying theses huge data has become difficult. Topic modeling is an unsupervised machine learning technique that can be used to get a summary from a large collection of documents. Topic modeling has been widely used in English content, yet the application of topic modeling in Arabic language is limited. Therefore, the aim of this paper is to provide a systematic review of the application of topic modeling algorithms in Arabic content. Using a well-known and trusted databases including ScienceDirect, IEEE Xplore, Springer Link, and Google Scholar. Considering the publication date from 2012 to 2022, we got 60 papers. After refining the papers based on predefined criteria, we resulted in 32 papers. Our result show that unfortunately the application of topic modeling techniques in Arabic content is limited.

Keywords:
Arabic content, Topic modeling, digital platform, LDA, topic classification.

1. Introduction

Topic modeling can be defined as the task to extract and detect topics from large collection of documents [1]. Topic modeling follow unsupervised machine learning to detecting topics [2]. Discovering topics is very important task in natural language processing (NLP) [1]. It can be used to cluster and classify text to benefit several fields such as healthcare, education, and E-Commerce [2]. As the amount of data on the web is getting larger by the day hence the mean of topic modeling is required to get the most advantage from this data [31]. Therefore, several topic modeling techniques have been proposed, the most popular one is Latent Dirichlet Allocation (LDA) which was discovered in 2003 [47]. Likewise, Non-negative matrix factorization (NMF) both are unsupervised technique which means they both operated on unlabeled data with a predetermined number of topics [2]. In addition, several approaches have been build based on LDA and NMF [1]. Thus, makes the interest toward topic modeling in English language is growing focus of research, Meanwhile, unfortunately, the application of topic modeling in Arabic language is limited [3].

Topic modeling has been used in several research disciplines [4]. Topic modeling applications in natural language processing (NLP) have gain the interest of many researchers in previous studies. For instance, software engineering can include large volume of unstructured data including use cases, source code, etc. Applying topic modeling to this data can improve many tasks in software engineering [5] .

Other applications involve social media data. Social media consider to be a vital source of information regarding many aspect of human life [6]. Analyzing this data can open up different opportunities to explore the real world [4]. Previous studies applied topic modeling to social medial data to mine public opinion regarding specific topic, analyze people expressions on social media platforms can benefit various fields such as marketing, public health, or political events, etc. [6]. As for Arabic language unfortunately the application of topic modeling is limited [3].

Arabic language is the language of the Holy Quran. It spread widely in Islamic countries. With more than 280 million native speakers, it considered as one of the most popular languages around the world [7]. In spite of its popularity, the applications of Arabic language in text mining are limited [8]. The reason for that is the complex structure of Arabic language, which makes Arabic language to be difficult to learn and analyze compared to other languages [3]. Therefore, the contribution of this paper is to provide an overview of the studies regarding the applications of topic modeling in Arabic language. The reminder of this paper is organized as following. Section 2 provides the detailed methodology of building the paper. Section 3 illustrates the result from analyzing the selected papers. Section 4 is the conclusion based on the reviewed papers.

Manuscript received June 5, 2023
Manuscript revised June 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.6.1

2. Methodology

We build a systematic review to get an overview of the application of topic modeling techniques in Arabic content starting from 2012 to 2022. To achieve this goal, we followed the review protocol illustrated in Fig. 1.

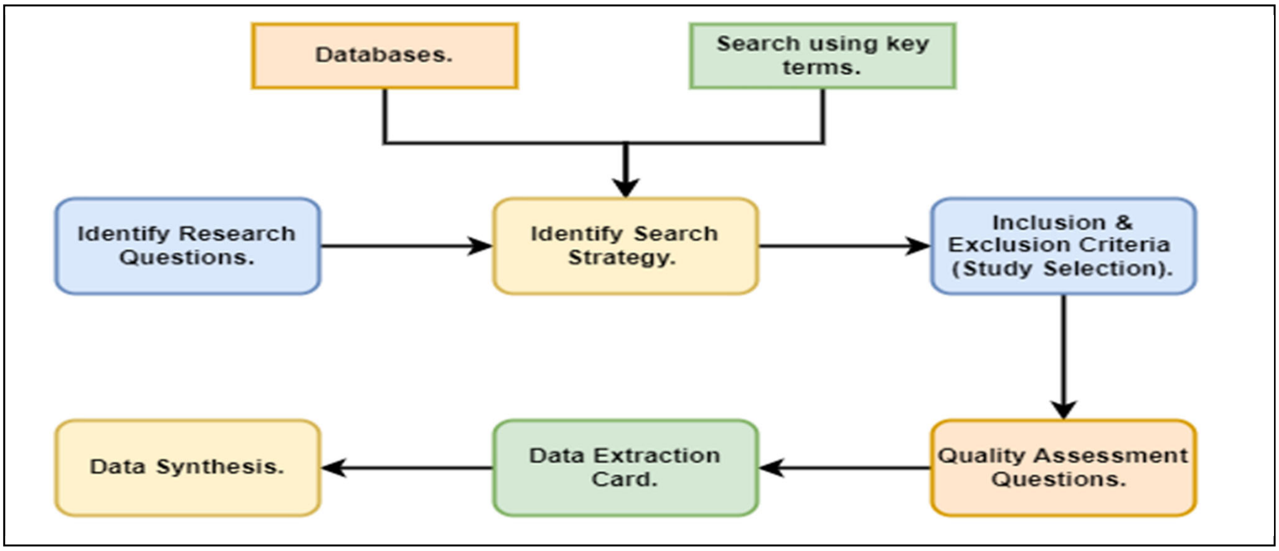

Databases.
Search using key terms.
Identify Research Questions.
Identify Search Strategy.
Inclusion & Exclusion Criteria (Study Selection).
Quality Assessment Questions.
Data Extraction Card.
Data Synthesis.

Figure 1. Review of protocol stages.

2.1 Research Questions

To achieve the purpose of this paper, we focus on three questions as stated in the following:

RQ 1: What are the applications of topic modeling in Arabic content?

This question assesses how topic modeling was used to benefit Arabic language.

RQ2: What are topic modeling Algorithms that have been used in the literatures?

Identifying topic modeling algorithms that was applied to detect topics from social media data.

RQ3: What are the sources of the data?

The aim of this question is to list sources that were used with topic modeling techniques. For new source that were not used with topic modeling yet, researchers can take the opportunity to explore them.

2.2 Search Strategy

This section will illustrate the search strategy to get the collected papers which involve identifying search terms, resources, and search process. To collect the papers, we used the different terms such as (Topic modeling Arabic) and (Twitter or social media or long document). The resources that have been used to extract the primary papers are obtained from four online databases which are Google scholar, IEEE Xplore, Springer Link, and ScienceDirect. The search was refined to obtain papers from 2012 to 2022. To get the candidate papers, a predefined set of key terms is used individually or combined to search the selected databases. Then the papers were downloaded in a folder for further selection process. After downloading the papers, we got 60 candidate papers. Fig. 2 illustrated how many papers were collected from each selected database**.**

2.3 Search Selection

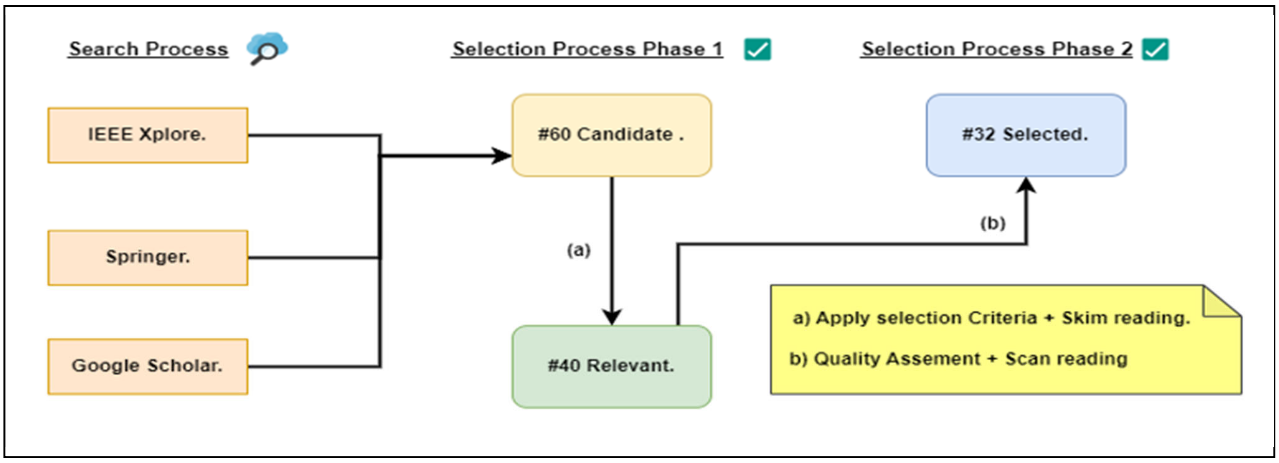
The result of the search process is 60 papers however, most of them are irrelevant to the objective of this paper or provide no answer to the research questions. Hence, we used the selecting strategy to find out the relevant papers from the candidate papers. The selecting strategy consists of the two following phases.

- Phase 1: Skim reading for 60 candidate papers and applying inclusion and exclusion criteria. After this phase we got 40 relevant papers.
- Phase 2: Scan reading to the relevant papers and apply quality assessment questions to test if the paper could answer one or more of the research questions. After phase two, we got 32 selected papers which represent primary studies. Fig. 2 shows in detail the process of each phase.

Table 1. List of Quality Assessment Questions.

No.	Quality Assessment Question.
1.	Does the paper contain clear tables and figures for the extracted topics?
2.	Is the purpose of the study clearly defined?
3.	Is the topic modeling algorithm clearly stated?
4.	Is the source of data collection clearly explained?

Figure 2. Search and Selection process.


Search Process
Selection Process Phase 1
Selection Process Phase 2
IEEE Xplore.
Springer.
Google Scholar.
#60 Candidate .
(a)
#40 Relevant.
(b)
#32 Selected.
a) Apply selection Criteria + Skim reading.
b) Quality Assement + Scan reading

2.4 Quality Assessment

To assess the quality of the collocated paper, we used several questions which are listed in table 1. The answer to these questions will be used to judge the paper in order to get the primary studies.

2.5 Data Extraction

To extract the primary data that will help us to answer research questions, we build the data extractor card as shown in table 2.

2.6 Data Synthesis

The purpose of this stage is to interrupt and analyze the data derived from the primary study. Results will be then represented and explained in figures and tables forms.

Table 2. The Data Extraction Card.

Year of Publication.
Publication type.
Title of the article.
RQ 1: What are the applications of topic modeling in Arabic content?
List of the applications of topic modeling in Arabic language.
RQ2: What are the topic modeling Algorithms?
Topic modeling algorithms that can detect topics from social media data.
RQ3: What is the source of the data?
Data collection that was used by topic modeling algorithms

3. Result

In the following subsections, we going to present the result of our systematic review.

3.1 Outline of The Selected Papers

At the start of our search, we got 62 papers published from 2012 to 2022. Papers are gathered from both journals and conferences. After refining them based on predefined criteria, we got 34 papers. Among the selected papers 59% were published in journals and 41% were published in conferences. The name of the journal and conference in addition to the distribution of the selected papers are shown in Table 3. All reviewed papers were summarized in Table 4. The summary includes the following description: the objective of each study, the publication year, topic modeling algorithm, and the dataset.

3.2 Applications of Topic Modeling in Arabic Content (RQ1).

Searching the literature, several studies have adopted different topic modeling algorithms to Arabic content. The following subsections highlighted the main related works of the applications of topic modeling in Arabic content. These applications include The Holy Quran, topic base sentiment analysis, and other applications.

3.2.1 The Holy Quran

Allah sent The Holy Quran to prophet Mohammed to guide every aspect of human life, it considered to be a robust source of information. However, this information is represented in unstructured format. Therefore, many tried to apply text mining techniques including topic modeling to help reader to get the semantic meaning from the holy Quran. The first study was by Abdul Sattar et al. [9], they applied LDA on each surah of the Quran to extract the most frequent terms and detect the main latent themes. Another application of LDA by Alhawarat [10]. The study was focused on only the Yusuf chapter of the holy Quran. The topic was extracted based on three shapes of word which are word, stem, and roots. Results shows that applying LDA is promising however combing LDA with other techniques can give butter results. Similar work has been done in [11], nonnegative matrix factorization algorithm (NMF) was applied to extract topics from a corpora of Quran's verses. The topics were visualized and linked with each verse. However, based on [12] the link between verses and the extracted topics was ambiguous.

3.2.2 Topic Based Sentiment Analysis.

With the age of technology, social media has become a huge part of our everyday activity. User generated content on social media is considered to be a vital source for several text mining techniques [3]. Sentiment analysis can be defined as the approach of discovering public opinion toward specific topic [13]**.** Combing topic modeling with sentiment analysis has been used widely in English content [14]. Unfortunately, its application in Arabic content is limited [3]. The main idea behind combining sentiment analysis with topic modeling is to discover the feeling toward a specific topic [15]. Studying the literature, few work has been published that adopted the topic-based sentiment analysis on Arabic content**.** Starting by the work in [16], which proposed an approach to discover sentimental events on a collection of Arabic tweets. They associated a topic with the emotion over a period of time. Result shows that the proposed approach gave promising result and open the door for any upcoming application.

Saidi et al. [17] proposed a novel approach to classify users account on twitter. The aim was to detect and reveal terrorist profile on twitter. The proposed framework was build based on Social Network Analysis (SNA) and Semi-Supervised Machine Learning (SSML) to detect the users based on their activities on twitter. BERT topic modeling was used to extract feature from the collected data. These features will be used as an input to sentiment analysis algorithm classifier to predict terrorist profiles.

In [18], topic-based sentiment analysis was used to classify user reviews and comments toward deferent topics. Their approach built around the assumption that words belong to the same topic have the same semantic. As the coverage ratio of semantic resources in Arabic language is limited comparing with other language, LDA was used to integrate the terms with the same semantic to different topics. Terms was linked with their concepts using BabelNet. Then Naïve Bayesian (NB), Decision Tree (DT), and

Table 3. Distribution and publication venues of the selected papers.

Venues name	***Journal***	***Conference***	***# of study***
Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences.		√	1
International Journal of Advanced Computer Science and Applications	√		1
Proceedings of SAI Intelligent Systems Conference		√	1
Complexity	√		1
International Journal of Business Intelligence and Data Mining	√		1
Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society		√	1
Innov. Smart Cities Appl.	√		1
Proceedings of the 2017 International Conference on Smart Digital Environment		√	1
The International Conference on Information, Communication & Cybersecurity.		√	1
Procedia Computer Science	√		2
IEEE Access	√		1
IET Networks	√		1
2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)		√	1
Journal of machine learning research	√		1
HAL	√		1
International Conference on Cloud Computing (ICCC).		√	1
17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim)		√	1
4th Conference on Data Mining and Optimization (DMO).		√	1
26th International Workshop on Database and Expert Systems Applications (DEXA).		√	1
2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA).		√	1
Proceedings of the 2018 International Conference on Signal Processing and Machine Learning.		√	1
Int. Arab J. Inf. Technol	√		1
Journal of computing and information technology	√		1
Procedia computer science	√		1
Engineering Applications of Artificial Intelligence	√		1
International Journal of Environmental Research and Public Health	√		1
Tenth International AAAI Conference on Web and Social Media.		√	1
International Journal of Advanced Computer Science and Applications	√		1
Arabian Journal for Science and Engineering	√		1
Journal of Medical Internet Research	√		1
Applied Intelligence	√		1
Proceedings of International Conference on Trends in Computational and Cognitive Engineering		√	1
Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 27	√		
Total	20	14	34
Percent	59 %	41 %	100 %

Support Vector Machine (SVM) were used for sentiment analysis. In addition, some studies tried to analyze the sentiments of a detected topic from the collected dataset. For instance, the studies [19-21] applied LDA with sentiment analysis on Moroccan colloquial and proved that topic-based sentiment is a promising field for any upcoming applications. Starting by the study [19], they used LDA and NMF algorithms to reveal the hidden topics from a collection of Moroccan tweets. Then they used Textblob to determine the sentiment of the detected topics. Finally, they compared the performance of LDA with NMF, result shows that the performance of LDA is better than NMF. Another application to Moroccan colloquial was performed in [20]. LDA was integrated with NB to associate each extracted topics with its sentiment. Hankar et al. [21] tried to analysis the sentiment of Morocco citizens after the break of the novel corona virus which known as (covid19). They applied LDA on a collected dataset composed of news comments of a popular Moroccan website called Hespress. Furthermore, they analyzed the sentiment of the extracted topics using SVM sentiment analysis technique. Finally, they stated that Moroccan citizen have negative feelings toward covid 19. All reviewed studies proposed promising approach for topic-based sentiment analysis, however due to the difficulty of Arabic colloquials, a lot of effort needs to be done for the preprocessing phase. Even though many studies have adopted topic-based sentiment analysis to detect and classify documents, there still a huge room for improvement. Table 5 summarized the algorithms that were used for both sentiment analysis and topic modeling in the reviewed papers.

3.2.3 Other Application

Topic modeling was also applied in deferent context. For instance, Fuad and Al-Yahya [22] tried to classify Arabic mobile apps using LDA. The study was carried out on Google play using the textual description of each apps. They started by using a crawler to collect data from Google play. Then they applied several natural language processing techniques to convert the unstructured forms of the data to a structured format ready for further analysis. Finally, they applied LDA to discover the latent themes and associate each app with the correct themes. Adel and Wang [24] tried to integrate LDA topic modeling with their proposed system, the aim behind this study was to detect terms that is difficult to be found on user tweet. They used twitter API to build their dataset. Then, they implemented the system using LDA algorithm to detect humanitarian crisis terms. Finally, they evaluated the proposed system and state that the proposed system can improve the classification process for Arabic tweets.

Alsaedi et al. [16] proposed a framework for event detection in Arabic tweets. They started by data collection then continue with several steps including data cleaning, feature selection, topic clustering, and summarization. The proposed framework was compared to LDA algorithm, and their result stated that the proposed framework performs better than LDA. Another application on twitter platform was in [17], their goal was to detect and reveal terrorist groups that aim to target innocent people. Their approach was based on BERTopic modeling. To evaluate the effect of topic modeling on clustering Arabic documents, Alghamdi et al. [26] conducted an experimental study on a large collection of newspapers websites. The study applied two clustering algorithms, Mean intra-cluster distance (MICD) and Davies-Bouldin index (DBI) then they

compared the result with Latent Semantic Analysis (LSA).

3.3 Topic Modeling Algorithms (RQ2).

Topic modeling approaches work under the assumption that topics can be detect from each document in the corpora [1]. This section describes the famous topic modeling algorithms that has been described in the reviewed papers**.**

3.3.1 Latent Dirichlet Allocation (LDA)

Blei et al. [22] proposed the popular topic modeling algorithm Latent Dirichlet Allocation (LDA). LDA is a generative model that detect topics from documents based on the statistical computation of words [23]. LDA assumed that each document is mixture of topics, and each topic is mixture of words. Moreover, each word in the topic is assigned to a probability based on the word occurrence [16]. Alomari et al. [39] used LDA topic modeling to detect topics related to public fears around covid 19 and the government measures to stop the viral spread of the virus. They applied LDA on a collection of twitter data, which is consist of 14 million tweets from Saudi

Arabia. Result shows that they were able to extract 15 government measures and six topics related to public concern around the pandemic.

To compare the performance of LDA with other topic modeling techniques, Abuzayed and Al-Khalifa

Table 4 list of all reviewed papers, purpose, year, topic modeling algorithm, and dataset.

Reference	Objective	publication	Algorithm	Dataset
[9]	Discover semantic meaning of the holy Quran	2013	LDA	Long document
[10]		2015	LDA	Long document
[11]		2014	NMF	Long document
[24]	Categorized Arabic mobile apps	2021	LDA	App description.
[25]	Events detection.	2020	Proposed approach.	Twitter dataset.
[18]	Classification of user reviews and comments	2019	LDA	Book Reviews.
[19]	Detect hidden topics from a collection of Moroccan tweets	2021	LDA, NMF	Twitter dataset.
[20]	Case study on Moroccan colloquial.	2017	LDA	Face book pages.
[21]	Effect of covid19 on Moroccan citizen	2022	LDA	News documents.
[26]	Detect term related to crises in Arabic language	2020	LDA	Twitter dataset
[16]	Event detection	2016	LDA based approach.	Twitter data
[27]	Studying how stemming effects the result of LDA.	2017	LDA.	
[28]	Comparison study.	2015.	PLSA and LSA	Web pages.
[29]	Experimental study.	2021.	BERT, LDA, and NMF	News documents.
[42]	Clustering Arabic documents	2018	LDA	News documents
[30]		2021.		Web pages.
[31]	An enhanced approach for topic modeling.	2015.	TFIDF	Collection of Arabic documents
[32]	Detecting topics from dark webpages.	2012	K-means	Web pages.
[33]	Topic identification in noisy Arabic text.	2015.	graph based approach (LIGA)	Arabic Noisy dataset
[34]			Proposed approach called Neural Text Categorizer (NTC)	ANTSIX dataset.

[35]	Proposed approach for topic identification.	2018	Proposed approach	Collection of Arabic articles corpus
[36]	Comparative study LDA and K-means.	2016.	LDA	Collection of Arabic documents.
[37]	Text classification based on LDA.	2012	LDA based approach.	
[38]	Topic modeling to enhance the performance of AdaBoost.MH algorithm.	2015	LDA	News document
[17]	Reveal terrorist profile on twitter.	2022.	BERT	Twitter
[39]	Detecting public concern and government measures after COVID 19	2021	LDA	
[40]	Classification of Arabic tweets	2022.	|BERT topic	
[41]	Determination of suspicious messages from Arabic tweets.	2020.	LDA	
[13]	Experimental study of LDA on Named Entity Recognition system in Arabic language.	2017	LDA.	Arabic Text corpora.
[15]	Detect the main topics on hate speech around covid 19	2020	NMF	Twitter data
[23]	Tracing the latest trends after the breakdown of covid 19	2021	LDA.	Facebook posts
[42]	Determination of the optimal number of topics.	2021	LDA	Long document

Table 5 Summarization of topic modeling algorithms and sentiment analysis techniques in the reviewed papers.

Reference	**Topic modeling algorithm**	**Sentiment analysis algorithm**
[18]	**LDA.**	NB, SVM and DT.
[19]	LDA, NMF	Textblob
[20]	LDA	NB

[21]		SVM
[17]	BERT	SVM and NB.

[29] conducted a comparison study to compare the performance of the BERTopic, LDA, and NMF topic modeling algorithms. The study was applied on a collected data set composed of 111,728 Arabic news documents. The results of experiential shows that BERT has better performance than LDA and NMF.

Some researchers suggested to combine topic modeling techniques with clustering algorithm to enhance the performance of document classification [30, 43]. For instance, Alhawarat and Hegazi [43] proposed a method to combine LDA with k-means to cluster Arabic documents. They started by applying LDA to a collection of a news dataset which is composed of 2700 documents of 9 category. Then they applied K-means to extract topics from the dataset. Based on the evaluation results, they stated that combing topic modeling with clustering algorithms enhanced the performance of the topic modeling. The work in [30] proved the same finding by applying the same approach to a dataset that has been collected from several website related to security issues. The work [36] conducted a comparative study between LDA and k-means. The study was applied on a large collection of Arabic documents to show that LDA has perform better than k-means.

3.3.2 Nonnegative Matrix Factorization (NMF).

Nonnegative matrix factorization is another statistical method that factorizes the corpora. If we assume that we have a matrix X, the NMF will approximate X to W and H matrix. In text mining, W will be a row that composed of words and H is the columns that composed of numbers [42]. In other word, W will represent the topics found on the document and H is the weight for each topic [42] .The produced matrix have non-negative coefficient [42]. Therefore, comparing to LDA, NMF considered to be able to produce more topics than LDA [29].

3.3.3 Other Proposed Approach

Due to complex internal structure and lack of standard corpus in Arabic language, some researchers proposed an approach for topic modeling. For instance, the work in [31] suggested to combine an enhanced stemming technique with Term Frequency/Inverse Document Frequency (TFIDF) topic modeling technique. Thus, the proposed approach will decrease the size of the corpus and enhance the process of topic extraction.

To identify topic from noisy text Abainia et. el. [31], proposed a graph-based approach called LIGA. After experimental result they extended the proposed approach in order to enhance the accuracy of topic identification. Another novel approach for topic identification in Arabic noisy text was in [32]. Their approach was called NTC, which stand for neural text categorizer. Evaluation result shows that the proposed approach has high performance.

Alsanad [35] proposed approach for topic identification based on discriminative multi nominal naïve Bayes (DMNB) classifier and frequency transform. The approach was implemented on three phases including preprocessing, features extraction and normalization, and text identification. Experimental result shows that the proposed approach perform more effective than traditional topic identification method.

Figure 3 illustrated the distribution of topic modeling algorithms on the selected papers. We can state that the most used topic modeling technique in the reviewed paper is LDA.

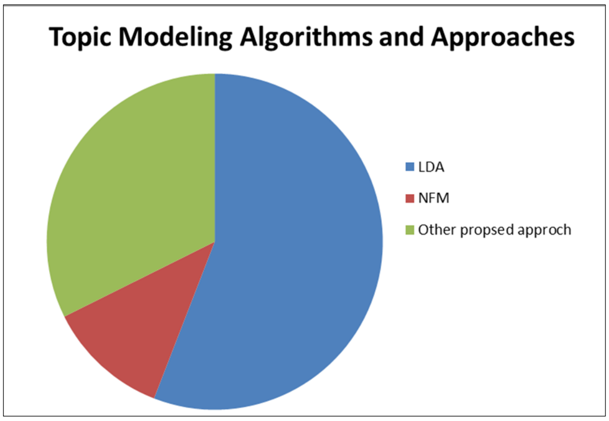


Figure 3. Distribution of topic modeling algorithms and approaches in the selected papers.

3.4 Data Source (RQ3)

Most of the used data in the reviewed publication is unstructured data. The examples of these data are social media post, news document, and more. Table 4 presented the dataset that has been used in the selected papers. As shown in the table the dataset included are mostly a social media post or long document such as news post. Each will be described briefly in the following subsections.

3.4.1 Social Media Posts

Social media is the language of the modern society many people turned to social media to express their experience with different aspect of their life. Thus, social media can be a robust source of information, many researchers tried to apply topic modeling algorithms on different social media platforms to benefit different sectors.
One of the most popular social media platforms is Twitter. Twitter is a free platform that enable user to write whatever they want in a message called tweet. These tweets can be feed to several text ming techniques to benefit several fields. Our review shows that most papers tend to feed twitter data to LDA algorithm or build an approach that was based on LDA to deal with this data. LDA has shown high performance working with short text such as tweet**.**

3.4.2 Long Document

Long document is unstructured form of data that has been collected from different source. These unstructured forms of data need to be transformed to a structured format that is ready to be fed to topic modeling algorithms. literatures used the Holy Quran, news document, or web pages as long document and applied several NLP techniques to transform it to a structured data. LDA proved its effectiveness to extract themes from theses long documents.

4. Conclusion

The aim of this paper is to present a comprehensive review of the recent research of the applications of topic modeling in Arabic content. From the reviewed papers, we can state that, due to its internal structure, Arabic language is considered to be a difficult language for many NLP techniques. In addition, there is a lack of Arabic standard datasets, standard dataset can facilitate the way for researchers to implement their research. Therefore, there are a limited number of research in the field of topic modeling in Arabic content, even though there are a lot of encouraging domains.

Another observation is that, even though several researchers have proposed novel techniques for topic modeling, we can state that the most used technique in the literature was LDA. Based on [44], LDA suffer from sparsity problem, hence using LDA with long documents was not a good choice and it may affect the process of topic extraction.
In the end, with all that has been done there are a huge room for improvement. Hence, Researchers should be encouraged to tackle and handle any challenges that face topic modeling in Arabic language.

References

[1] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems,* vol. 94, p. 101582, 2020.

[2] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2017: IEEE, pp. 745-750.

[3] S. Basabain, "A survey of Arabic thematic sentiment analysis based on topic modeling," *International*

Journal of Computer Science & Network Security, vol. 21, no. 9, pp. 155-162, 2021.

[4] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications,* vol. 78, pp. 15169-15211, 2019.

[5] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 34, no. 3, pp. 1427-1445, 2020.

[6] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.(IJACSA),* vol. 6, no. 1, 2015.

[7] M. Alhawarat, M. Hegazi, and A. Hilal, "Processing the text of the Holy Quran: a text mining study," *International Journal of Advanced Computer Science and Applications,* vol. 6, no. 2, pp. 262-267, 2015.

[8] A. Rafea and N. A. GabAllah, "Topic detection approaches in identifying topics and events from Arabic corpora," *Procedia computer science,* vol. 142, pp. 270-277, 2018.

[9] M. A. Siddiqui, S. M. Faraz, and S. A. Sattar, "Discovering the thematic structure of the Quran using probabilistic topic model," in *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 2013: IEEE, pp. 234-239.

[10] M. Alhawarat, "Extracting topics from the holy Quran using generative models," *International Journal of Advanced Computer Science and Applications,* vol. 6, no. 12, pp. 288-294, 2015.

[11] M. H. Panju, "Statistical extraction and visualization of topics in the qur'an corpus," *Student. Math. Uwaterloo. Ca,* 2014.

[12] M. Alshammeri, E. Atwell, and M. A. Alsalka, "Quranic topic modelling using paragraph vectors," in *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*, 2021: Springer, pp. 218-230.

[13] I. El Bazi and N. Laachfoubi, "Arabic named entity recognition using topic modeling," *context,* vol. 230, 2017.

[14] B. Bansal and S. Srivastava, "On predicting elections with hybrid topic based sentiment analysis of tweets," *Procedia Computer Science,* vol. 135, pp. 346-353, 2018.

[15] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of hate speech in covid-19–related tweets in the arab region: Deep learning and topic modeling approach," *Journal of Medical Internet Research,* vol. 22, no. 12, p. e22609, 2020.

[16] N. Alsaedi, P. Burnap, and O. Rana, "Sensing real-world events using Arabic Twitter posts," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2016, vol. 10, no. 1, pp. 515-518.

[17] F. Saidi, Z. Trabelsi, and E. Thangaraj, "A novel framework for semantic classification of cyber terrorist communities on Twitter," *Engineering Applications of Artificial Intelligence,* vol. 115, p. 105271, 2022.

[18] M. Bekkali and A. Lachkar, "Arabic sentiment analysis based on topic modeling," in *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*, 2019, pp. 1-6.

[19] N. Habbat, H. Anoun, and L. Hassouni, "Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets," in *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, 2021: Springer, pp. 147-161.

[20] T. Zarra, R. Chiheb, R. Moumen, R. Faizi, and A. E. Afia, "Topic and sentiment model applied to the colloquial Arabic: a case study of Maghrebi Arabic," in *Proceedings of the 2017 international conference on smart digital environment*, 2017, pp. 174-181.

[21] M. Hankar, M. Birjali, A. El-Ansari, and A. Beni-Hssane, "Arabic Topic Modeling-Based Sentiment Analysis on COVID-19 Feedback Comments," in *Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21*, 2022: Springer, pp. 87-95.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, no. Jan, pp. 993-1022, 2003.

[23] A. Amara, M. A. Hadj Taieb, and M. Ben Aouicha, "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis," *Applied Intelligence,* vol. 51, pp. 3052-3073, 2021.

[24] A. Fuad and M. Al-Yahya, "Analysis and classification of mobile apps using topic modeling: A case study on Google Play Arabic apps," *Complexity,* vol. 2021, pp. 1-12, 2021.

[25] M. Daoud and D. Daoud, "Sentimental event detection from Arabic tweets," *International Journal of Business Intelligence and Data Mining,* vol. 17, no. 4, pp. 471-492, 2020.

[26] G. Adel and Y. Wang, "Detecting and Classifying Humanitarian Crisis in Arabic Tweets," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2020: IEEE, pp. 269-274.

[27] R. Baly *et al.*, "Comparative evaluation of sentiment analysis methods across Arabic dialects," *Procedia Computer Science,* vol. 117, pp. 266-273, 2017.

[28] H. Alghamdi and A. Selamat, "Topic modelling used to improve Arabic web pages clustering," in *2015 International Conference on Cloud Computing (ICCC)*, 2015: IEEE, pp. 1-6.

[29] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Computer Science,* vol. 189, pp. 191-194, 2021.

[30] A. R. Alharbi, M. Hijji, and A. Aljaedi, "Enhancing topic clustering for Arabic security news based on k-means and topic modelling," *IET Networks,* vol. 10, no. 6, pp. 278-294, 2021.

[31] A. Alsaad and M. Abbod, "Enhanced topic identification algorithm for Arabic Corpora," in *2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim)*, 2015: IEEE, pp. 90-94.

[32] H. M. Alghamdi and A. Selamat, "Topic detections in Arabic dark websites using improved vector space model," in *2012 4th Conference on Data Mining and Optimization (DMO)*, 2012: IEEE, pp. 6-12.

[33] K. Abainia, S. Ouamour, and H. Sayoud, "Topic Identification of Noisy Arabic Texts Using Graph Approaches," in *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 2015: IEEE, pp. 254-258.

[34] K. Abainia, S. Ouamour, and H. Sayoud, "Neural Text Categorizer for topic identification of noisy Arabic Texts," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, 2015: IEEE, pp. 1-8.

[35] A. Alsanad, "Arabic Topic Detection Using Discriminative Multi nominal Naïve Bayes and Frequency Transforms," in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 2018, pp. 17-21.

[36] A. Kelaiaia and H. F. Merouani, "Clustering with probabilistic topic models on arabic texts: a comparative study of LDA and K-means," Int. Arab J. Inf. Technol., vol. 13, no. 2, pp. 332-338, 2016.

[37] M. Zrigui, R. Ayadi, M. Mars, and M. Maraoui, "Arabic text classification framework based on latent dirichlet allocation," Journal of computing and information technology, vol. 20, no. 2, pp. 125-140, 2012.

[38] B. Al-Salemi, M. J. Ab Aziz, and S. A. Noah, "LDA-AdaBoost. MH: Accelerated AdaBoost. MH based on latent Dirichlet allocation for text categorization," Journal of Information Science, vol. 41, no. 1, pp. 27-40, 2015.

[39] E. Alomari, I. Katib, A. Albeshri, and R. Mehmood, "COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning," International Journal of Environmental Research and Public Health, vol. 18, no. 1, p. 282, 2021.

[40] M. Hernandez-Mendoza, A. Aguilera, I. Dongo, J. Cornejo-Lupa, and Y. Cardinale, "Credibility Analysis on Twitter Considering Topic Detection," Applied Sciences, vol. 12, no. 18, p. 9081, 2022.

[41] M. A. AlGhamdi and M. A. Khan, "Intelligent analysis of Arabic tweets for detection of suspicious messages," Arabian Journal for Science and Engineering, vol. 45, pp. 6021-6032, 2020.

[42] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, "Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)," in Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, 2021: Springer, pp. 341-354.

[43] M. Alhawarat and M. Hegazi, "Revisiting k-means and topic modeling, a comparison study to cluster arabic documents," IEEE Access, vol. 6, pp. 42740-42749, 2018.

[44] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 1445-1456.