Comparative Analysis of Topic Modeling Algorithms for Short Texts in Persian Tweets

Amir Hossein Karimi[1*], Masoud Akbari[1] and Mohammad Akbari[1]

[1]The Computer Science Department, Amirkabir University of Technology, Tehran, Iran.

*Corresponding author(s). E-mail(s): ahkarimi@aut.ac.ir;
Contributing authors: ma.akbari421@aut.ac.ir;
akbari.ma@aut.ac.ir;

Abstract

Topic modeling is a popular natural language processing technique to uncover hidden patterns and topics in extensive text collections. However, there is a lack of comprehensive studies that focus specifically on applying topic modeling algorithms to short texts, particularly from social media platforms. Even fewer studies have explored comparing different topic modeling algorithms for low-resource languages such as Persian. Our study aims to address this gap by thoroughly investigating topic modeling algorithms and metrics tailored for short texts, particularly Persian tweets. We collected and preprocessed a substantial dataset of Persian tweets. We also developed a dedicated tool that enables reproducibility and facilitates the evaluation of various topic modeling algorithms using the provided dataset. Our comparative analysis included Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Indexing (LSI), Gibbs Sampling Dirichlet Mixture Model (GSDMM), and Correlated Topic Model (CTM). To measure their performance, we employed well-established metrics, namely Purity, Normalized Mutual Information (NMI), and Coherence. Our experimental results indicate that GSDMM and CTM+BERT exhibit superior performance compared to other algorithms in terms of purity and NMI on the Persian short-text topic modeling dataset. Additionally, CTM+BERT demonstrates competitive coherence performance compared to GSDMM.

Our study provides valuable insights into the effectiveness of different topic modeling approaches for short texts and can help researchers select the most appropriate algorithm for their specific use case.

Keywords: Topic modeling, Social Media, Short texts, Persian language

1 Introduction

Social media platforms have become an integral part of modern society, offering users a variety of benefits that reach far beyond just entertainment [1, 2]. In social media, there is a wealth of information that can be analyzed and used to gain a better understanding of society's opinions, beliefs, and attitudes. The analysis of these topics over time can provide valuable insights for decision-makers, shedding light on the intellectual and ideological perspectives of individuals and society [3–5], also traditionally known as topic modeling. While topic modeling is a well-established task in natural language processing, centered on the identification of underlying themes or topics in large text corpora [6–9], the available methods for conventional texts produce unsatisfactory results for short text, i.e., social media data. Topic modeling techniques applied to social media data present notable differences from conventional approaches used for topic modeling, particularly those applied to large documents. These dissimilarities can be categorized as follows:

- **Data structure:** Social media datasets tend to be more susceptible to noise and may contain irrelevant information. Conversely, traditional topic modeling commonly relies on curated resources that maintain exact topic-related content [10]. Moreover, papers collectively suggest that shortness and lack of co-occurrences can negatively impact the ability to model topics accurately. [11–13]
- **Well-defined topics:** Conventional topic modeling approaches are typically employed on collections of texts where topics are clearly defined and explicit. However, social media data frequently includes texts that implicitly refer to a topic, making it more challenging to elicit and decipher the underlying themes. [14].
- **Behavior dependency:** The diverse behavior exhibited by users on social media platforms results in diverse data forms. In contrast, conventional topic modeling assumes datasets that conform to identified regularities and patterns [15].

Considering the widespread impact of social media platforms like Twitter and Instagram in our daily lives, which serve as platforms for diverse opinions and information shared by users, there has been a notable surge in interest within the research community towards topic modeling of short texts [16]. Despite the extensive research conducted in the field of topic modeling, there is still a significant dearth of comprehensive reviews that thoroughly assess

the effectiveness of both classic and contemporary methods for topic modeling in the context of short-text analysis, social media data, and low-resource languages, such as Persian. Consequently, the task of selecting the optimal method for these languages poses a considerable challenge, as each topic modeling approach brings forth its own set of strengths and weaknesses [17].

This study aims at investigating topic modeling on social media data for low-resource languages, concentrating primarily on the Persian language. In this regard, we have preprocessed a corpus of Twitter data in Persian to create a standard dataset for topic modeling. In order to demonstrate the effectiveness of cutting-edge topic modeling algorithms for low-resource languages, we compared their performance on this dataset and reported our findings in a comprehensive way.

The objectives and contributions of this paper can be summarized as follows:

- The preparation of a topic modeling dataset consisting of Persian text posts from social networks.
- The application of appropriate preprocessing and normalization techniques to ensure the quality and consistency of the prepared data.
- The implementation of widely adopted topic modeling techniques for the analysis of Persian text data.
- The comprehensive evaluation of the performance of these methods on a Persian text dataset containing various topics.
- The comparison and assessment of the results generated from the different methods, with the aim of identifying the most effective approach for topic modeling of Persian text data.

2 Literature Review

The primary objective of this study is to collect a new dataset for topic modeling in the low-resource Persian language. The studies have been divided into two categories to facilitate a clear understanding of the relevant literature. The first category includes studies on collecting new datasets for topic modeling and the various approaches used. The second category includes studies addressing the difficulties and opportunities in topic modeling in low-resource languages.

2.1 Datasets

Given the limitless availability of ever-expanding information, data collection will be an essential component of every study. A common way of collecting topic modeling datasets is to gather data from reliable resources. These kinds of resources may have a pre-categorization of their documents, resulting in a cleaner dataset such as 20 Newsgroups [18]. Conversely, the data relayed on social media does not have a pre-defined category. It is a typical practice

for researchers to utilize hashtags to gather data from social networks [19–22]. Chen et al. [21] utilized the hashtag *#engineeringProblems* to gather a dataset and study the casual talks of engineering students on Twitter in an effort to better understand the challenges that engineering students face in their education. Using Twitter tweets as a reference point, Egger et al. [20] evaluate the performance of several algorithms in terms of their strengths and flaws from a social science perspective. They utilized *#covidtravel* and the combination of *#covid* and *#travel* to get tweets for data collection.

Keywords are another alternative for data collection in social media. For instance, Athukorala and Mohotti [23] utilized two datasets concerning *'Organic Food'* and *'COVID-19'* that were collected using the keywords *'Covid'* and *'Organic Foods.'*

Similarly, Dahal et al. [24] analyzed climate change tweets, collecting tweets containing specific keywords relating to climate change. Kim et al. [25] aim to identify consumers' preferences and perceptions of genderless fashion trends and study to evaluate consumers' awareness of the current gender-less fashion trend using the text-mining method. They examined posts that included the keyword *'genderless fashion'* in social media.

There are further data collection methods. For instance, using geotag locations in addition to using hashtags or keywords [21, 24]. Furthermore, special methods have been used by certain scholars, such as the usage of emojis to gather tweets [19]. Excellent progress has been made, and a wide range of methodologies has been used, yet certain areas remain unexplored. The specific hashtags and data-gathering techniques utilized in certain studies were not disclosed. The quality of the final data set may be greatly impacted by the method used to choose the appropriate hashtag; therefore, this choice should not be ignored.

2.2 Low-resource languages

Researchers have a tendency to believe that the behaviors of English-language users in social networks are representative of those of users of other languages. However, studies indicate that social network users from various languages utilize Twitter for different preferences and purposes [26]. Researchers have addressed the challenges associated with these languages in the realm of topic modeling for low-resource languages by proposing innovative approaches. Daben Liu et al. [27] introduced a semi-supervised model for topic classification, which, unlike unsupervised topic modeling algorithms, requires labeled data for accurate topic classification. Although reliant on labeled data, their method has proven to be suitable for low-resource languages. The authors validated the effectiveness of their approach by successfully applying it to Malay language data. However, developing and evaluating topic modeling methods for low-resource languages pose additional challenges. Evaluating the coherence of topics in such languages can be problematic due to the scarcity of resources and the need for external data. To address this issue, Shudong

Hao et al. [28] proposed a multilingual method specifically designed to overcome the challenges of evaluating topic coherence in low-resource languages. Furthermore, Ray et al. [29] conducted a comprehensive review and evaluation of various topic modeling methods, including LSI, NMF, and LDA, on Hindi texts, which is another low-resource language. Their study contributes to understanding the performance and applicability of different topic modeling techniques in the context of Hindi. Several studies have been conducted on low-resource languages, showcasing diverse approaches and applications in the field. For instance, G"uven et al. [30] utilized NMF and LDA algorithms to evaluate sentiment in Turkish tweets. Additionally, Habbat et al. [31] focused on analyzing Moroccan tweets to extract relevant data, identify distinct moods, and illustrate frequently occurring subjects. Moreover, numerous researchers have collected and studied multilingual social network datasets, enriching the resources available for studying low-resource languages [28, 32].

To the best of our knowledge, there is currently no publicly available topic modeling dataset specifically in the Persian language. This gap in available resources has motivated researchers like Hosseini et al. to concentrate on measuring and tracking the evolution of the pandemic response using Persian tweets [33]. In a similar vein, Parvin Ahmadi et al. proposed a method for Persian text classification, wherein topic models were utilized as a means of classifying Persian texts [34].

Despite the significant presence of Persian language users on social media platforms, previous studies exploring large-scale topic modeling in Persian social networks have been relatively scarce. This study aims to address this research gap by focusing specifically on the Persian language, which has been largely overlooked in previous attempts to model large-scale topics in the Persian social network.

3 Data Collection and Preprocessing

Twitter is one of the most widely used social media platforms, making it a popular source for collecting data. There are two primary methods for gathering data from social media: *using APIs* or *crawling*. The former is a straightforward and commonly used approach that involves collecting data based on keywords, hashtags, users, and other criteria through Twitter's API. However, this method has limitations, such as being restricted to the last 3200 users' posts and the number of requests that can be sent within a specific period (approximately 900 requests in 15-minute intervals) [35].

The second method, crawling, offers an alternative approach to collecting data with fewer restrictions. In this study, we have used crawling to obtain data from Twitter and developed an open-source tool based on this approach that can be publicly accessed[1]. The following section describes the steps taken to collect topic modeling data in more detail.

[1] https://github.com/DSInCenter/topicmodel

3.1 Choosing the topics

It is important to use trustworthy sources when gathering news. One approach is to collect tweets from official news agencies, which can provide reliable data. However, this method has both advantages and disadvantages. While it results in a standardized dataset using the official language, it is limited to the topics covered by news agencies and cannot capture all the subjects that people tweet about [36, 37]. As a result, all Persian Twitter accounts were used to collect tweets without placing any restrictions on user accounts. The data collection process is visually presented in Figure 1

Generally speaking, two major approaches exist for gathering tweets: keywords and hashtags [19–21, 25]. While keywords help gather a set of tweets related to a particular topic, they may also result in collecting many unrelated tweets. For instance, in business and economy, one of the keywords used is *'cheap'*. However, some tweets containing this keyword may be irrelevant to the topic, such as the tweet *"Don't you know where to have a cheap time bomb??? My phone alarm won't wake me up anymore!"*. Hashtags were utilized instead of keywords to guarantee a relevant data set. However, the proper selection of hashtags is essential for collecting a good dataset. Hashtags were chosen in a two-stage process. Initially, the top 100 frequently used hashtags were extracted from a temporary dataset containing all users' tweets within one month[2]. In the second step, appropriate hashtags for each topic were manually selected to produce the final dataset, and the dataset was subsequently crawled again. The table containing the hashtags used to gather the final tweets can be found in Table 1 [3].

3.2 Data exploration

This section includes several analyses aimed at improving our understanding of the dataset. The insights gathered from these analyses provide valuable information about the data's characteristics.

3.2.1 Data preprocessing

Before conducting topic modeling, it is essential to preprocess the dataset. This involves eliminating all non-Persian characters, except for spaces, commas, and question/exclamation marks, as they can reveal themes. Arabic characters are substituted with their Persian counterparts. Duplicate data is also eliminated, and tweets containing less than five tokens or twenty characters are excluded to ensure a clearly defined topic.[4].

[2]When the dataset was created, it covered a period of one month between 2022-02-10 and 2022-03-10

[3]The dataset primarily consists of crawled Persian content. However, for the purpose of inclusivity, this article has translated the hashtags and other keywords into English, allowing non-Persian speakers to comprehend them.

[4]The code to reproduce the dataset with all these preprocessing steps can be found on the paper's GitHub page.

Table 1 Topics and their related hashtags that are used to collect data

Topics	Hashtags
Business and Economy	stock, oil, employment, economic, economic_news, iran_economy, free_marker, central_bank, currency, tax inflation, rate, sanction, gold, 4200_currency expensiveness, bank, car, forex, petrol market, currency_rate, euro, oil_price, budget price, economy, coin, OTC, share share, insurance
Medical and Health	corona, ministry_of_health, no_to_vaccination, vaccine yes_to_vaccination, omicron, medical, doctor, mask forcible_vaccination, health, i_will_vaccinate, corona_outbreak, covid19, minister_of_health, Barkat best_vaccine_is_the_most_accessible
Sports	esteghlal, perspolise, football, sports, halamadrid, real_madrid, sports_is_not_political, bartar_league, government_team, taj, arsenal, pirouzi, tractor, farhad_majidi, volleyball, olympics, hamed_lak, fair_football, derby, FIFA, Liverpool, penalty, Fenerbahçe, league, blue_federation, political_sports, Chelsea, RealPSG, WorldCup, mehdi_taremi, team, tennis, club
Art	poem, book, cinema, theater, movie, series, music, movie_offer, Hafez, Saadi, book_offer, director, singer, fajr_movie_festival, drama, moviestar, movietime
Technology	internet, apple, Samsung, game, google, bitcoin, crypto, Ethereum, digital_currency, BTC, starlink, hamrah_aval, SEO, Irancell, Cardano, digikala, smart,
Transportation	congestion, snapp, taxi, airplane, metro, bus, train, congestion_plan, airport, provinancial_trip, mehrabad_airport, chaloos_road
Educations	teacher, education, university, konkoor, azad_university, educational_testing, tehran_university, education_ministry, student, teachers, teacher_day, schools, farhangian, farhangian_university
Religion	Allah, Imam, Ramezan, Quran, Islam, Hajj
Lifestyle	beauty, cooking_schedule, skin, cooking, mod, fruits, food, cafe, restaurant
Social	rain, weather, water, earth_quake, drought, mars_8th, suicide, marriage, divorce, poverty, men, child_abuse, political_prisoner, women_rights
Environment	women_world_day, woman, women, family, hijab, pollution, air_pollution, dust_storm, tehran_weather blizzard, dust, rainfall, flood, raining, tehran_air_pollution, crisis_management, fire, environment, sand, meteorology, meteorology_twitter

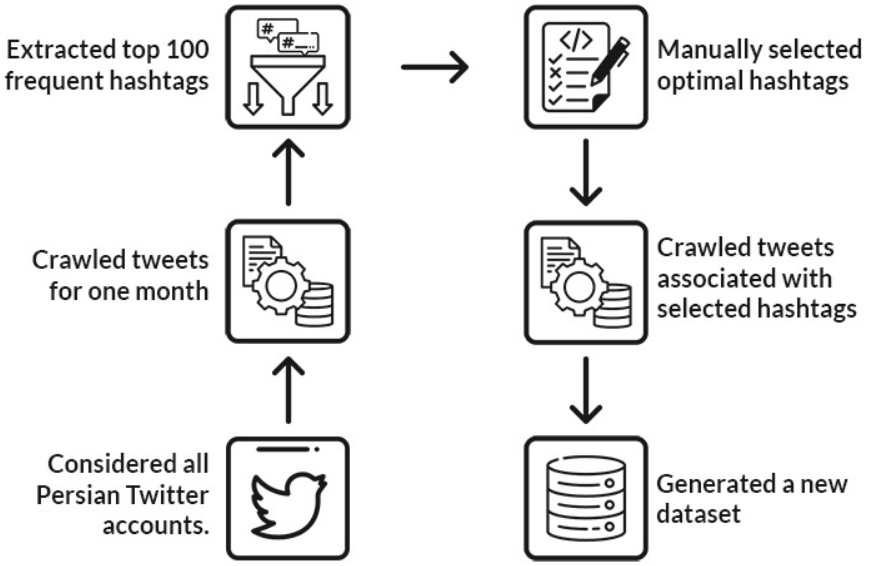

Fig. 1 Data collection process for Persian short text topic modeling dataset

3.2.2 Statistics of the dataset

Table 1 outlines the division of the dataset into 11 distinct topics for further analysis and organization. Additionally, each tweet is accompanied by metadata that can be utilized for future projects. Table 2 presents six key attributes for each topic, including the number of tweets, users, hashtags, average tweet length, tokens, and unique words. Identifying these characteristics can help interpret data and provide useful insights.

3.2.3 Distribution of topics

The dataset is evenly distributed across all topics, with 10,000 tweets in each topic. Initially, a larger collection of tweets was gathered, but it caused issues while conducting experiments and slowed down the model's performance. As a result, in order to preserve balance, the frequency of each topic was limited to 10,000 tweets.

3.2.4 Most frequent words

Each topic in the dataset has its own collection of salient words that represent the topic to which it belongs. In order to better understand and visualize these common words, word clouds have been generated for each topic. These word clouds, shown in Figure 2, represent the most common words for each topic. Additionally, Table 3 provides a list of the five most frequent words for each

Table 2 Key attributes linked to each topic

Topics	Number of tweets	Number of users	Number of hashtags	mean-length of tweets	Number of tokens	Number of unique words
Business and Economy	10000	3449	26668	109	176884	24521
Medical and Health	10000	4159	22797	110	157481	24965
Sports	10000	3557	22371	103	162724	25090
Art	10000	4695	17638	74	128904	23553
Technology	10000	3262	25057	89	146661	21167
Transportation	10000	6217	28784	108	157402	27888
Education	10000	3987	34224	119	171696	31230
Religion	10000	3957	17333	92	133242	24115
Lifestyle	10000	4746	54626	109	170571	31230
Social	10000	4797	30942	117	182388	30911
Environment	10000	4543	14389	104	165245	26507

Table 3 Five most frequent words of each topic; numbers in tuples show the frequency of each word of the topic

Topics	Five most frequent words
Business and Economy	(stock, 4660), (dollar, 1953), (market, 1803),(price, 1707), (Iran, 1601)
Medical and Health	(Corona, 7443), (vaccine, 5882), (Omicron, 1953), (Iran, 1663), (person, 1566)
Sports	(Persepolis, 5638), (Esteghlal, 3854), (team, 2989), (football, 2226), (match, 1975)
Art	(Hafez, 3665), (film, 3459), (Saadi, 3326), (book, 1424), (heart, 1329)
Technology	(bitcoin, 3637), (bit, 2187), (coin, 2082), (dollar, 1488), (crypto, 1335)
Transportation	(Snapp, 5430), (congestion, 2350), (Tehran, 1998), (subway, 1996), (driver, 1650)
Educations	(teachers, 7843), (global, 4091), (protests, 3049), (gathering, 2804), (teacher, 2781)
Religion	(Imam, 2227), (Allah, 997), (Islam, 670), (Appearance, 669), (Friday, 644)
Lifestyle	(clothes, 617), (food, 378), (skin, 263), (Iran, 260), (mod, 182)
Social	(women, 4072), (woman, 2782), (poverty, 2545), (Iran, 2143), (hijab, 1963)
Environment	(water, 3522), (earthquake, 2573), (weather, 2415), (pollution, 2273), (environment, 1860)

topic, allowing for a more detailed examination of the vocabulary associated with each topic.

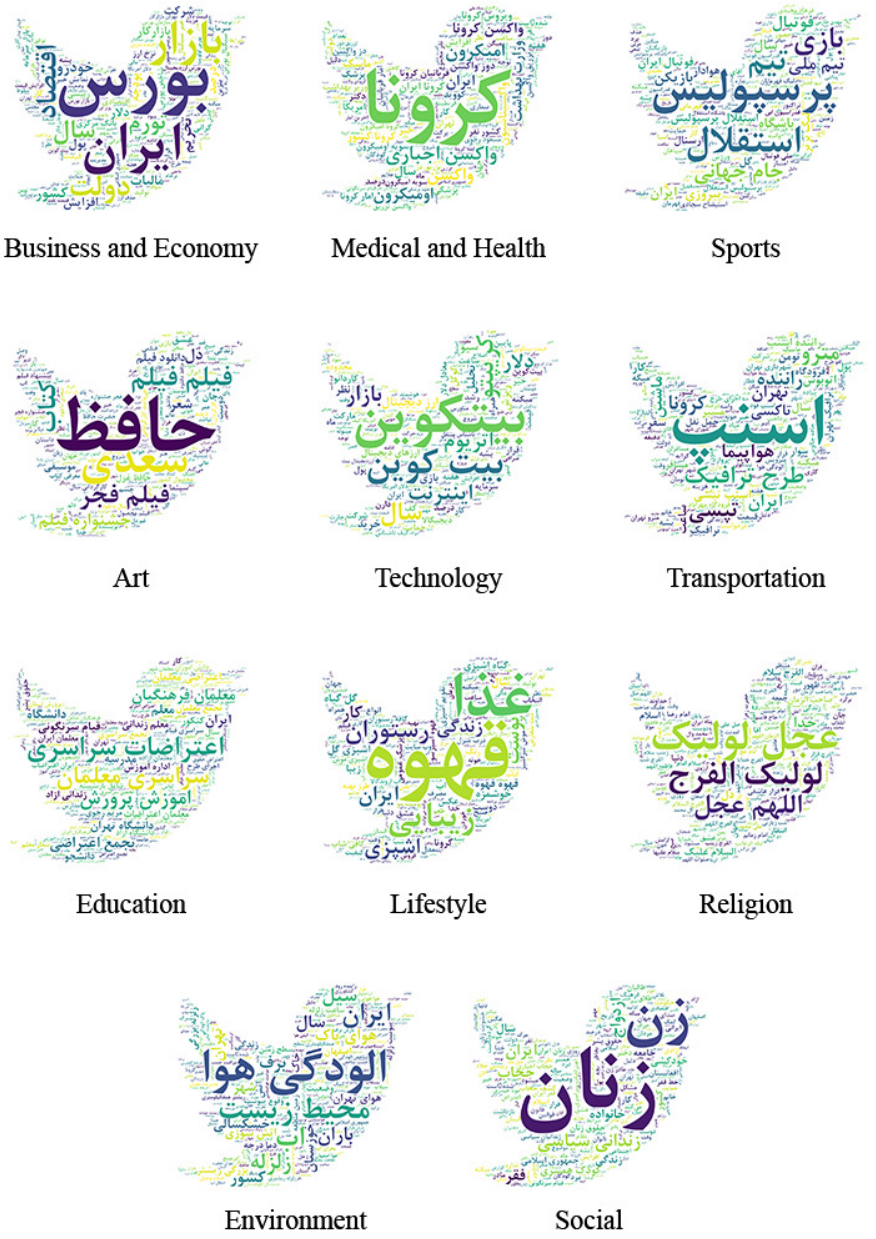

Fig. 2 The word clouds of each topic

3.2.5 Important hashtags

Hashtags play an essential role in categorizing and discovering tweets. They make tweets more searchable and help in locating relevant topics. Table 4 provides the top five hashtags associated with each topic, demonstrating their semantic proximity. This suggests that the hashtag-based dataset collection method has been fairly accurate.

Table 4 Five most frequent hashtags of each topic; numbers in tuples show the frequency of each hashtag of the topic

Topics	Five most frequent hashtags
Business and Economy	(stock, 3341), (dollar, 821), (market, 795), (price, 665), (Iran, 416)
Medical and Health	(Corona, 4228), (vaccine, 1406), (no_to_compulsory_vaccination, 904), (Omicron, 757), (Iran, 536)
Sports	(Perspolis, 3087), (2860, 3854), (football, 904), (worldcup, 776), (Iran, 346)
Art	(Hafez, 3402), (Saadi, 2430), (fajr_movie_festival, 669), (book, 626), (movie, 510)
Technology	(bitcoin, 1861), (BTC, 1749), (crypto_currency, 1106), (crypto, 1033), (ethereum, 721)
Transportation	(Snapp, 3937), (subway, 1317), (Tapsi, 1260), (airplane, 905), (traffic, 753)
Educations	(teachers, 4426), (teacher, 1196), (Iran, 1082), (university, 797), (education, 666)
Religion	(Allah, 739), (Islam, 446), (Quran, 214), (Imam-Reza, 201), (Imam, 121)
Lifestyle	(coffee, 2666), (beauty, 2027), (food, 1632), (cook, 1485), (restaurant, 871)
Social	(women, 2120), (poverty, 1896), (hijab, 1042), (woman, 970), (engagement, 892)
Environment	(water, 1738), (earthquake, 1681), (pollution, 1372), (environment, 1176), (snow, 716)

4 Experiments

The primary objective of this article is to make available a Persian dataset containing short texts. Accordingly, a selection of classic and state-of-the-art methods will be employed to establish a baseline for future applications of the dataset. This section will explain the methods utilized and the metrics employed.

4.1 Methodology

In the realm of topic modeling, analyzing short texts poses unique challenges compared to long documents, such as limited contextual information, single topic focus, and semantic dispersion [38, 39]. To address these challenges, various short-text topic modeling methods have been developed. However, there

is a lack of systematic comparison between different models. In this article, we have selected methods suitable for both lengthy and short-text modeling to assess their performance on our collected dataset.

Latent Dirichlet Analysis (LDA). LDA [6] is a well-known topic modeling method that is typically used with long texts or documents. It assumes that each document contains a set of words, and each word is associated with a single topic. In addition, each document is composed of multiple topics with varying proportions, which can assist in identifying the main topic of the document. The first step in this method is to construct a dictionary of all unique words across the documents. Then, each topic is represented by a distribution of a set of words, with words that are strongly related to the topic having a higher probability in the distribution. It is essential to specify the number of topics in the initial step.

Non-Negative Matrix Factorization (NMF). NMF [40] is a method initially used for feature extraction that can reduce high-dimensional vectors to a lower-dimensional latent space. This technique produces two matrices, one indicating the topics and the other indicating the significance of words within each topic.

Dirichlet Multinomial Mixture Model. The Gibbs Sampling algorithm for Dirichlet Multinomial Mixture (GSDMM) [41] is a topic modeling method designed specifically for short texts. GSDMM addresses the challenge of sparsity in short text topic modeling and presents word topics, much like LDA. The key difference between GSDMM and LDA is that GSDMM assumes that a tweet or text contains only one topic, while LDA assumes that a document can contain multiple topics.

Biterm Topic Modeling (BTM). BTM [13] is a topic modeling method that considers documents as a set of word pairs, unlike other methods that focus on individual words. Firstly, a dictionary of word pairs is created, and then a distribution of word pairs is calculated in a similar process to that of LDA.

Structural Topic Modeling (STM). STM [42] is a generative method for topic modeling method that utilizes the probabilities of words belonging to each topic and the frequency of their occurrences, considering each document as a blend of various topics. Unlike other methods with a similar approach, STM takes advantage of metadata for each document. For instance, in the case of topic modeling the dataset in question, hashtags can be employed as metadata.

Topic Modeling of Word Embeddings (TMWE). Word embeddings are widely used in various text and natural language processing tasks, including topic modeling. In TMWE [9], short texts are transformed into vectors in a latent space, and the negative log-likelihood is then minimized to generate a set of topics for the documents.

ProdLDA. The computation of the posterior probability is a complex task and can be time-consuming. Any slight change in the dataset's probability distribution will require the posterior probability to be recomputed. To address these issues, the ProdLDA [43] method has been developed, which employs a Variational Autoencoder. The autoencoder maps each document to a latent probability distribution, resulting in faster computation times than LDA and eliminating recomputation.

Latend Semantic Indexing (LSI). LSI [8] is also another topic modeling method that employs matrix factorization. This approach utilizes the factorization of non-singular values. LSI assumes that words sharing the same context within a topic have similar meanings. Its name, latent semantic indexing, stems from its ability to establish relationships between words and phrases in a latent space.

Contextualized Topic Modeling (CTM). Many traditional topic modeling methods are unable to handle words that are not included in the training corpus, which can make it challenging to apply these methods to new languages or data. CTM [7] offers a solution to this problem by using contextualized word embeddings instead of Bag-of-Words. This allows CTM to handle unknown words and support additional languages more effectively. For example, the mean of 300-dimensional Persian-Fasttext [44] word embeddings can be used to represent tweets in Topic Modeling of Word Embeddings. Additionally, pretrained versions of Persian-AlBert [45] and Persian-BERT [45] with the CLS token can be used as embedding layers in CTM.

4.2 Metrics

Three distinct evaluation metrics were utilized to assess the efficacy of each method on the gathered dataset from diverse perspectives. While some of these metrics are typically employed for evaluating clustering approaches, given that topic modeling is an unsupervised task akin to clustering [6], they are also relevant for this purpose. The chosen metrics are outlined below.

Purity. The purity metric first assigns a label to each cluster by selecting the most frequent ground-truth label. The accuracy of this assignment is then assessed by dividing the number of documents correctly assigned to their respective cluster by the total number of documents in that cluster. It is important to note that increasing the number of clusters could potentially result in higher purity scores [46]. To better understand the model's performance, an additional metric should be utilized in conjunction with purity.

Normalized Mutual Information (NMI). NMI measures the extent of entropy reduction of labels in a cluster. NMI values range from 0 to 1, with higher values indicating clusters with lower entropy [46]. Since NMI is normalized during computation, it facilitates comparison between clusterings with different numbers of clusters.

Coherence. Let $T = \{w_1, w_2, ..., w_n\}$ be a topic derived by a topic model and represented by the top-n most likely terms. The greater the average pairwise similarity between words in T, the more coherent the topic [47].

A predicted label was required to apply evaluations, so each of the 11 chosen topics was mapped to one of the clusters generated by the topic modeling algorithm. The mapping process involved selecting each cluster's most frequent ground truth label.

5 Results and Analysis

Our objective is to investigate the most suitable topic modeling algorithms for various NLP tasks. To achieve this end, we will analyze algorithms from various perspectives. Ultimately, we will present the findings as a decision tree diagram (Figure 3) that can assist in selecting an appropriate topic modeling method.

Comparative Analysis. The results of the conducted experiments are presented in Table 5. It is important to note that the relatively lower performance of LDA on the dataset was expected due to its inherent suitability for longer documents. Conversely, GSDMM, an enhanced variant of LDA designed specifically for shorter texts, demonstrated superior performance. It is noteworthy that both LDA and GSDMM belong to the category of generative algorithms and are categorized as probabilistic graphical models.

When considering selecting a topic modeling method, if the focus is on capturing complex relations between topics, CTM or ProdLDA are recommended approaches. CTM incorporates neural networks to incorporate contextual information, enabling it to capture intricate connections between topics and their context. It excels at modeling fine-grained relationships and dependencies among topics within a context. Similarly, ProdLDA combines topic modeling with deep neural networks to effectively capture complex relationships between topics.

The results indicate that CTM and ProdLDA demonstrate high purity, indicating that most tweets are accurately assigned to the most dominant topic. In contrast, LDA, LSI, or BTM algorithms may not explicitly address complex relations between topics. LSI, for instance, relies on a large corpus of text data to capture latent semantic relationships between terms and documents. The poorer performance of LSI could be attributed to its assumption of orthogonal basis vectors, which restricts its ability to identify topics accurately.

Insights for Method Selection. Based on these results we provide following suggestion for selecting the best topic modeling algorithm for the dataset at hand.

(1) Computation cost is a crucial factor to consider when selecting a topic modeling algorithm. Methods like CTM or TMWE heavily rely on word embeddings and often require constructing or loading pre-trained word embedding models. In our study, with a moderate-sized dataset, we efficiently

Table 5 The results of applying selected models on the dataset

Methods	Purity	NMI	Coherence
LDA	0.30	0.15	0.14
ProdLDA	0.54	0.39	**0.67**
NMF	0.53	0.41	0.52
BTM	0.40	0.32	0.46
STM	0.47	0.38	0.49
LSI	0.42	0.19	0.40
TMWE	0.47	0.38	0.49
CTM + ALBERT	0.48	0.36	0.61
CTM + BERT	0.56	0.41	0.65
GSDMM	**0.58**	**0.46**	0.53

obtained the embedding vectors for the CTM algorithm using GPUs. However, it is important to note that computation requirements can significantly increase for larger datasets, necessitating careful consideration.

In contrast, LDA involves estimating topic-word and document-topic distributions, which can be accomplished using standard probabilistic inference techniques. This typically results in a more manageable computation cost compared to methods that rely on word embeddings. Additionally, BTM, which focuses on pairwise co-occurrence modeling, generally exhibits lower computation costs than other algorithms.

(2) The presence or absence of metadata can significantly impact the choice of topic modeling algorithm in certain situations. Metadata considerations should be taken into account when selecting an appropriate algorithm.

CTM, ProdLDA, GSDMM, and STM are examples of supervised or semi-supervised algorithms that may require labeled data for training or fine-tuning. These algorithms might rely on additional metadata, such as author information or keyword annotations, to guide the topic modeling process. By incorporating some form of supervision, utilizing labeled data or metadata, these algorithms aim to enhance the results of topic modeling.

On the other hand, LSI, TMWE, BTM, NMF, and LDA are unsupervised algorithms. They do not rely on labeled training data or explicit topic annotations. These algorithms automatically infer topics from the input text without additional metadata, making them suitable in cases where metadata is unavailable or not a primary consideration.

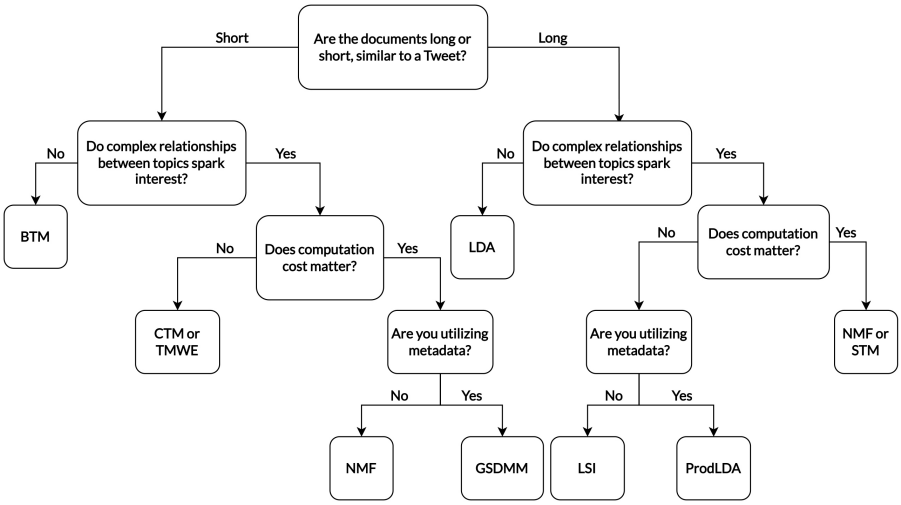

Fig. 3 The Diagram Decision Tree for Method Selection

6 Conclusion and Future Work

In summary, this paper makes a valuable contribution to the field of topic modeling by providing a publicly available dataset of short text from Persian Twitter. Additionally, a dedicated tool was developed and shared with the research community, enabling the replication of results and the exploration of various topic modeling algorithms on the dataset. The study conducted extensive research on topic modeling algorithms and metrics, yielding important insights into the effectiveness of different approaches for short texts, specifically tweets. The findings revealed that GSDMM, an enhanced version of LDA designed for short-length texts, outperformed LDA and other methods. Moreover, CTM+BERT demonstrated competitive performance and exhibited better coherence in comparison to GSDMM.

Future studies could extend this research by investigating the effectiveness of topic modeling algorithms on other languages and short-text datasets. Additionally, there is potential for the development of new algorithms and metrics aimed at improving the performance of topic modeling on short texts. Furthermore, the publicly available dataset and tool provided in this study offer opportunities for other researchers to explore different research questions within this domain.

References

[1] Fuchs, C.: Social media: A critical introduction (2014). https://doi.org/10.4135/9781446270066

[2] van Dijck, J.: The Culture of Connectivity: A Critical History of Social Media (2013). https://doi.org/10.1093/acprof:oso/9780199970773.

001.0001

[3] Tufekci, Z., Wilson, C.: Social media and the decision to participate in political protest: Observations from tahrir square. Journal of communication **62**(2), 363–379 (2012). https://doi.org/10.1111/j.1460-2466.2012.01629.x

[4] Hargittai, E.: Digital na (t) ives? variation in internet skills and uses among members of the "net generation". Sociological inquiry **80**(1), 92–113 (2010). https://doi.org/10.1111/j.1475-682X.2009.00317.x

[5] Li, Z., Fan, Y., Jiang, B., Lei, T., Liu, W.: A survey on sentiment analysis and opinion mining for social multimedia. Multimedia Tools and Applications **78**, 6939–6967 (2019). https://doi.org/10.1007/s11042-018-6445-z

[6] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[7] Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E.: Cross-lingual contextualized topic models with zero-shot learning, 1676–1683 (2021). https://doi.org/10.18653/v1/2021.eacl-main.143

[8] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6), 391–407 (1990). https://doi.org/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9

[9] Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020). https://doi.org/10.1162/tacl_a_00325

[10] Hu, B., Song, Z., Ester, M.: User features and social networks for topic modeling in online social media. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 202–209 (2012). IEEE

[11] Pedrosa, G., Pita, M., Bicalho, P., Lacerda, A., Pappa, G.L.: Topic modeling for short texts with co-occurrence frequency-based expansion. In: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 277–282 (2016). https://doi.org/10.1109/BRACIS.2016.058

[12] Qiang, J., Chen, P., Wang, T., Wu, X.: Topic modeling over short texts by incorporating word embeddings. In: Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21, pp. 363–374 (2017). Springer

[13] Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456 (2013)

[14] Vosecky, J., Jiang, D., Leung, K.W.-T., Xing, K., Ng, W.: Integrating social and auxiliary semantics for multifaceted topic modeling in twitter. ACM Transactions on Internet Technology (TOIT) **14**(4), 1–24 (2014)

[15] Pu, X., Wu, G., Yuan, C.: User-aware topic modeling of online reviews. Multimedia Systems **25**, 59–69 (2019)

[16] Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter, 80–88 (2010). https://doi.org/10.1145/1964858.1964870

[17] Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012). https://doi.org/10.1145/2133806.2133826

[18] Lang, K.: Newsweeder: Learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339 (1995)

[19] Bhatnagar, S., Choubey, N.: Making sense of tweets using sentiment analysis on closely related topics. Social Network Analysis and Mining **11**(1), 1–11 (2021). https://doi.org/10.1007/s13278-021-00752-0

[20] Egger, R., Yu, J.: A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in Sociology **7** (2022). https://doi.org/10.3389/fsoc.2022.88649

[21] Chen, X., Vorvoreanu, M., Madhavan, K.: Mining social media data for understanding students' learning experiences. IEEE Transactions on learning technologies **7**(3), 246–259 (2014). https://doi.org/10.1109/TLT.2013.2296520

[22] Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys, 241–249 (2010)

[23] Athukorala, S., Mohotti, W.: An effective short-text topic modelling with neighbourhood assistance-driven nmf in twitter. Social Network Analysis and Mining **12**(1), 1–15 (2022). https://doi.org/10.1007/s13278-022-00898-5

[24] Dahal, B., Kumar, S.A., Li, Z.: Topic modeling and sentiment analysis of global climate change tweets. Social network analysis and mining **9**(1), 1–20 (2019). https://doi.org/10.1007/s13278-019-0568-8

[25] Kim, H., Cho, I., Park, M.: Analyzing genderless fashion trends of consumers' perceptions on social media: using unstructured big data analysis through latent dirichlet allocation-based topic modeling. Fashion and Textiles **9**(1), 1–21 (2022). https://doi.org/10.1186/s40691-021-00281-6

[26] Hong, L., Convertino, G., Chi, E.: Language matters in twitter: A large scale study. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, pp. 518–521 (2011). https://doi.org/10.1609/icwsm.v5i1.14184

[27] Liu, D., McVeety, S., Prasad, R., Natarajan, P.: Semi-supervised topic classification for low resource languages. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5093–5096 (2008). https://doi.org/10.1109/ICASSP.2008.4518804

[28] Hao, S., Boyd-Graber, J., Paul, M.J.: Lessons from the Bible on modern topics: Low-resource multilingual topic model evaluation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1090–1100. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1099

[29] Ray, S.K., Ahmad, A., Kumar, C.A.: Review and implementation of topic modeling in hindi. Applied Artificial Intelligence **33**(11), 979–1007 (2019). https://doi.org/10.1080/08839514.2019.1661576

[30] Güven, Z.A., Diri, B., Çakaloğlu, T.: Comparison method for emotion detection of twitter users. In: 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–5 (2019). https://doi.org/10.1109/ASYU48272.2019.8946435. IEEE

[31] Habbat, N., Anoun, H., Hassouni, L.: Topic modeling and sentiment analysis with lda and nmf on moroccan tweets. In: The Proceedings of the Third International Conference on Smart City Applications, pp. 147–161 (2021). https://doi.org/10.1007/978-3-030-66840-2_12. Springer

[32] Lopez, C.E., Gallemore, C.: An augmented multilingual twitter dataset for studying the covid-19 infodemic. Social Network Analysis and Mining **11**(1), 1–14 (2021). https://doi.org/10.1007/s13278-021-00825-0

[33] Hosseini, P., Hosseini, P., Broniatowski, D.A.: Content analysis of persian/farsi tweets during covid-19 pandemic in iran using nlp (2020). https://doi.org/10.18653/v1/2020.nlpcovid19-2.26

[34] Ahmadi, P., Tabandeh, M., Gholampour, I.: Persian text classification based on topic models. In: 2016 24th Iranian Conference on Electrical Engineering (ICEE), pp. 86–91 (2016). IEEE

[35] Rate limits, docs, twitter developer platform. Twitter (2023). https://developer.twitter.com/en/docs/twitter-api/rate-limits Accessed 2023-03-02

[36] Jin, Z., Cao, J., Zhang, Y., Luo, J.: News verification by exploiting conflicting social viewpoints in microblogs. Proceedings of the AAAI Conference on Artificial Intelligence **30**(1) (2016). https://doi.org/10.1609/aaai.v30i1.10382

[37] Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**(2), 211–36 (2017). https://doi.org/10.1257/jep.31.2.211

[38] Qiang, J., Qian, Z., Li, Y., Yuan, Y., Wu, X.: Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering **34**(3), 1427–1445 (2020)

[39] Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 937–946 (2009)

[40] Zhao, R., Tan, V.Y.: Online nonnegative matrix factorization with outliers. IEEE Transactions on Signal Processing **65**(3), 555–570 (2016). https://doi.org/10.1109/TSP.2016.2620967

[41] Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering, 233–242 (2014). https://doi.org/10.1145/2623330.2623715

[42] Roberts, M.E., Stewart, B.M., Tingley, D.: stm: An r package for structural topic models. Journal of Statistical Software **91**(2), 1–40 (2019). https://doi.org/10.18637/jss.v091.i02

[43] Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: International Conference on Learning Representations (2017)

[44] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893 (2018)

[45] Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M.: Parsbert: Transformer-based model for persian language understanding. Neural Processing Letters **53**, 3831–3847 (2021)

[46] Schutze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval (2008)

[47] Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics, Los Angeles, California (2010). https://aclanthology.org/N10-1012