Journal of Intelligent & Fuzzy Systems 41 (2021) 2441–2457
DOI:10.3233/JIFS-202545
IOS Press

Topic prediction and knowledge discovery based on integrated topic modeling and deep neural networks approaches

Zeinab Shahbazi and Yung-Cheol Byun*
Department of Computer Engineering, IIST, Jeju National University, Jeju Special Self-Governing Province, Korea

Abstract. Understanding the real-world short texts become an essential task in the recent research area. The document deduction analysis and latent coherent topic named as the important aspect of this process. Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) are suggested to model huge information and documents. This type of contexts' main problem is the information limitation, words relationship, sparsity, and knowledge extraction. The knowledge discovery and machine learning techniques integrated with topic modeling were proposed to overcome this issue. The knowledge discovery was applied based on the hidden information extraction to increase the suitable dataset for further analysis. The integration of machine learning techniques, Artificial Neural Network (ANN) and Long Short-Term (LSTM) are applied to anticipate topic movements. LSTM layers are fed with latent topic distribution learned from the pre-trained Latent Dirichlet Allocation (LDA) model. We demonstrate general information from different techniques applied in short text topic modeling. We proposed three categories based on Dirichlet multinomial mixture, global word co-occurrences, and self-aggregation using representative design and analysis of all categories' performance in different tasks. Finally, the proposed system evaluates with state-of-art methods on real-world datasets, comprises them with long document topic modeling algorithms, and creates a classification framework that considers further knowledge and represents it in the machine learning pipeline.

Keywords: Machine Learning, knowledge discovery, Topic Modeling, Latent Dirichlet Allocation, Short Text, Long Short Term Memory

1. Introduction

The field of Machine Learning has evolved over the last decades, and the development in this field is proposed in [1, 2]. This process is a famous area in topic modeling because of containing various solutions to extract information from short texts. Furthermore, classical domains, exceptional deep learning become very important in different practical science and engineering [3–5]. ML algorithms prompt researchers to permit driving cars using the computer, writing and publishing reports related to sport, etc. These algorithms operate to capture the knowledge from the selected data. Another advantage of this system is that it doesn't need programming; instead, it is accomplished with the system based on process improvement and alteration. It learns the operation steps to use based on data, which is one of the procedure's issues manually. The technology of traditional databases is limited to reading, writing, querying, and other operations, but it is not available to extract knowledge out of data.

*Corresponding author. Yung-Cheol Byun, Department of Computer Engineering, IIST, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Korea. E-mail: ycb@jejunu.ac.kr.

Knowledge discovery in a database can extract information from structured and unstructured data, e.g., XML, text file, etc. [6]. Terminated knowledge is supposed to be in the format of machine-readable and machine-interpretable and should propose the knowledge interface. In different fields, data collection needs a dramatic speed. Significant communities like government corporates and scientific communications are overcome with the rush data types typically found in online database systems. Extracting useful and meaningful part for analyzing this type of data is unmanageable without using powerful tools. Applying statistics and systematic packages can make filters and interpret output results even though it has advantages of applying it correctly [7]. The immediate and essential need for a new generation that uses computational tools and techniques to extract knowledge from required data is the subject of knowledge discovery surface in the database (KDD) [8, 9]. Traditional topic modeling algorithms, e.g., latent Dirichlet allocation (LDA) [10], and probabilistic latent semantic analysis (PLSA) [11] is used to discover information from document contents without the need for prior annotation or document labeling. Different ranges and documents are applied in the presented algorithms and contain various texts and topics with various words for each topic. Mentioned algorithms had a significant impact in other modeling collections, e.g., news articles, blogs, and research papers [12, 13]. Knowledge-based techniques are applied to a massive area of research and science. Using this method captures the search results and particular information through knowledge domains and knowledge classification, it is also effective in search strategies and text retrieval system design [14]. Regardless of knowledge-base method utility, it focuses on system performance, which only executes the planned process [15]. A simple description of knowledge discovery is the process of recognizing valid, novel, potentially useful, and eventually realize hidden parts of data [16]. Generally, it determines the data subset definition. A considerable part of the data definition is data processing, which frequently demonstrates knowledge discovery experimentation, iteration, user interaction, and various design decision and customizations. Identifying advanced knowledge requires the ways to extract knowledge from the dataset. The easy process of knowledge extraction can be complicated and challenging, but the final result can be considerable and compensate [17]. Major knowledge discovery systems require tools to extract detailed information

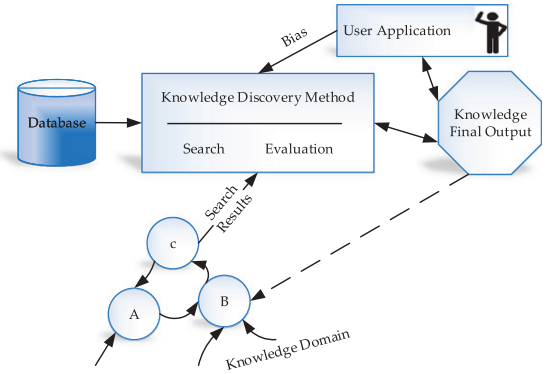

Fig. 1. Knowledge Discovery Framework in Database.

from an unstructured social media dataset. One of the tools generally applied in information extraction from social media content is an ontology used to find the concept and relationship in domain knowledge. Figure 1 shows the simple architecture of the knowledge discovery framework. Based on the available database, the knowledge discovery process is divided into search and evaluation sections. The search results are the clustering outputs, which are categorized into various domains. The clustering methods in this system are based on conceptual clustering and traditional numeric strategies, which maximize and produce the highest similarity within classes. The evaluation section is the process for qualifying the extracted knowledge from the dataset. The user applications applied for further processing of this system. Based on the defined process and conditions, the proposed system's results need to identify human beings that evaluate and identify the useful clusters. The knowledge discovery process input is the collected raw data from the internet sources and databases, data dictionary, and additional knowledge domains defined by the user for high-level focus. This process's output is the discovered knowledge sent back to the system as a new knowledge domain. The final feedback of this process saves into the knowledge domain for further processing the system.

The general goal of knowledge discovery based on topic modeling using a machine learning interface is to develop the system to extract the hidden knowledge of social media contents and reach the level of usable knowledge. The remaining of this paper is categorizing into five sections, which are the following: Section 2 explains the literature review about the combination of machine learning and topic modeling based on the majority of state-of-art topic models. Section 3 gives detailed information on the

system methodology, design architecture, and overall transaction process of the proposed system. Section 4 elaborates on the hybrid approach. Section 5 explains the implementations, and section 6 shows the proposed method's results. We are finally concluding the paper in section 7.

2. Literature review

In this section, we explain the related works. Section 2.1 presents the associated results on machine learning, Section 2.2 presents related works on LDA topic modeling, and section 2.3 presents the related work on knowledge discovery.

2.1. Machine learning

Machine learning is one of the most significant knowledge discovery areas with various famous algorithms for data processing, knowledge extraction, and learning behavior improvement. This process is to discover related information or hidden knowledge from the dataset, which is not available for humans. In entire research areas related to computer science, machine learning has the fastest growth in different technical fields. It also has various application domains, e.g., smart city, smart garden, smart factory, etc., directly related to daily human life, e.g., recommender system, voice recognition, etc. Data management is one of the machine learning achievements, containing the database, scientific analysis, statistical analysis, and expert systems. Short text topic modeling has various research directions in the machine learning field. Visualization, evaluation, checking the model, and deep learning listed for this direction. The visualization shows the topics based on the most repeated words in each category. In this case, more useful information presented by topic modeling related to document structure, which is helpful to extract the important parts of the document [18, 19]. The main problem of topic modeling is the evaluation, which is the challenge of this field for researchers [20]. The topic coherence is different for each topic, which new evaluation metrics required. Model-checking in topic modeling is based on the results of dataset performance. One solution for checking the model is through interactive visualization based on interpretive hypotheses. Finally, the development of deep learning techniques in this area gives the ability to automatically learning systems based on low dimensional representations. Autoencoder [21], document neural autoregressive distribution estimator [22], etc. The combination of topic modeling and deep learning techniques used in recent research topics too [23]. Combining deep learning techniques with topic modeling contains some benefits for exploring future research direction in the deep knowledge domain. Albalawi et. al. [24], focused on applying the supervised machine learning techniques to overcome the automatic text classification problems. The supervised text classification categorizes the documents into various predefined classes based on their subjects. The core part shows that users are able to extract information from textual information based on various patterns. Ayoub et. al. [25], proposed system estimated the deep focus on the similarity of the document by applying the K-Nearest Neighbour algorithm. The similarity process is to classify the sentiments toward neutral polarity based on mSMTP measure.

2.2. LDA topic modeling

During the past few years, there has been a lot of research published on topic modeling based on neural networks, i.e., Boltzmann machines and softmax layers [26–37]. A recurrent neural network (RNN) is also used to gather dynamic relationships in data. In topic modeling, topics are designed using RNN [38]. The neural network is also used for embedding learned based on NLP. A word embedding is represented as a high dimensional vector of words, which is learned from data. It enables relatedness of a word which the related words to another term as the summation of terms, e.g., $woman + king - man = queen$. Vivek Kumar et al. [18] proposed the soft clustering on the short texts based on the low-dimensional word2vec technique and Gaussian models for objective and subjective evaluation of documents. Similarly, a well known neural network-based word embedding approach is the word2vec [39]. Likewise, lda2vec and word2vec (W2V) are recent topic modeling based on word embedding for learning word embedding and LDA topics. These models predict the words in the document, word embedding, and also topic distribution. Usually, topic modeling is a vast area to automatically discover hidden thematic information from a text document with meaningful content [40–43]. Another aspect of topic modeling is to consider purification of the document clustering by unsupervised machine learning strategy, which means opposite to document. In the topic modeling procedure, many topics can occur in

the individual document, but frequent topics have more training set processes. Document clustering is the process of finding the similarity between documents and categorize them into meaningful groups. A good clustering system is a type of cluster which have incredibly deal with document characteristic. In recent years, graph-based clustering (spectral clustering) [44], which focuses on partitioning graphs, is one of the popular topics in the document clustering area. The proposed model defines the given document as an undirected graph, and each node is presenting a text document. Weight shows the edge of the document and returns the similarity between contents. There are two types of clustering methods. Hierarchical and k-means algorithms. The first one contains the single link and groups them based on the ward's method and the second one is for providing the information. Ximing et. al. [45] proposed the recently developed technique for aggregating the short texts into pseudo-documents. Self-Aggregation based topic modeling (SATM) process the shots texts without the need for heuristic information. To approach the fast interface mini-batch scheme presented and similarly, the Latent Topic Modeling (LTM) was applied to consider the short texts as standard input, but the text memberships were initially unknown.

2.3. Knowledge discovery

Knowledge discovery description is a practical interdisciplinary which processes various fields [46–49]. The majority of relevant areas are statistics, machine learning, artificial intelligence and reasoning with uncertainty, databases, knowledge acquisition, pattern recognition, information retrieval, visualization, intelligent agents for distributed and multimedia environments, digital libraries, and management information systems [50]. This field's major challenge is understanding the document content and deciding on uncertain documents, which is a lot in social media resources and probabilistic due to extremely impressed learning statistics and artificial intelligence. Complex data types need enough solutions and required content for knowledge discovery. The opposite probability permits anonymous data to predict the final decision. Some real-world examples are mentioned to make knowledge discovery more understandable: Academic research models are one complex topic that provides the activity sequences to extract the information from them. Another type is the hybrid models, which developed on the basis of the CRISP-DM. Johannes et al. [51] applied text mining techniques based on various dimensional analysis for the multi-dimensional knowledge representation (MKR). This technique processed into English and German documents. The analysis output of this system contains the sentiment relationship of documents, topic detection, etc. Table 1 representing the process of knowledge discovery on the selected dataset. Some sections justify more explanation. Step four is critical and can have more details. Indeed, most of the cases need to solve search and cataloging problem before the verified subsequent analysis. This process may provide a special requirement to overcome issues. In classical pattern-recognition, it is famous for feature extraction issues. The domain knowledge is required to overcome the problems.

3. Algorithmic approach used in methodology

This section represents the proposed algorithmic approaches applied in the proposed system. Section 3.1 presents related strategies to LDA topic modeling, and section 3.2 describes the LSTM associated works.

3.1. Language modeling

Function learning is based on language modeling which evaluate the probability of an activity log $q(S|model)$, or a sentence as Equation 1:

$$S = (S_1, S_2, ..., S_n) \tag{1}$$

Similarly, the same function can apply in the prediction of the subsequent activity or word. Other techniques that use the same system are listed as, e.g., LDA, which utilizes the bag-of-words strategy. Although, RNN modeling also use this model to exclude loss of temporal information to model log as shown in Equation 2.

$$q(S_n|S_1, S_2, ..., S_{n-1}, model) \tag{2}$$

3.1.1. Latent Dirichlet allocation

There are various defined statistic models in machine learning and natural language processing that one of the popular statistic models is LDA. LDA algorithm is unsupervised learning, and it's one of the machine learning algorithms toolboxes. The main concept of LDA is determined to find the similarity between documents and categorize the contents

Table 1
Involved phases related to knowledge discovery process

ID	Steps	Samples
1	Application domain comprehension, related knowledge and aim of user	By focusing on recent technology we can catch the related information
2	Selecting the data type which need to process on knowledge discovery system	Involve data consideration
3	Pre-process and cleaning steps on selected dataset	Applying fundamental performance on data noises and extract needed contents
4	Transformation and data reduction	Searching for effective information
5	Using data mining	To determine the aim of process

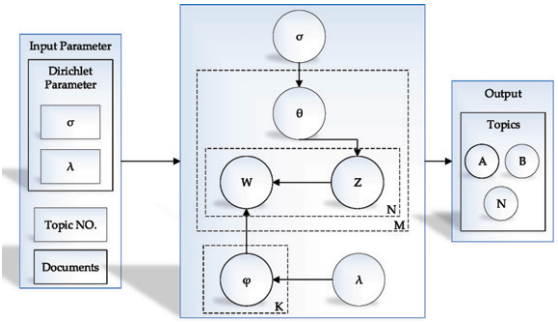

Fig. 2. Plate Notation of Latent Dirichlet Allocation (LDA).

into different parts, named as topics. It is a generative probabilistic model that consider extracting the hidden part of documents based on the conditional distribution. Depend on the applied dataset; each topic contains the same meaning contents, e.g., sport, news, education, etc. This procedure is repeated for all datasets. Figure 2 shows the process of the LDA technique in topic modeling. This system's main functionality is to categorize the provided dataset based on its meaning in various groups. The input parameters are fed into LDA system, and after preprocessing, the system cluster the similarity between them into various topics. This process repeats for the whole of the collected dataset.

3.2. Long Short-Term Memory (LSTM)

There are some issues in context learning related to traditional topic modeling. Temporal aspects are essential for any language model. Thus, to perform this task, an appropriate process is required. An RNN type, LSTM, can efficiently acquire knowledge of temporal features and context and further classify large time-series datasets.

The main goal of the RNNs network is to generate recurrent neural network connections to memorize. In various applications, language models that are based on RNN have recently determined the performance of state-of-art. Recurrent neural network dynamic behavior makes them desirable for sequential classification based issues. To train the input data in the first step, the word A system input and next-word, and the output are the systems following further actions. This process is continuously training on the whole dataset.

As a simple explanation of the RNN procedure, we can define more detail in terms of (Σ, C, δ) where inputs are defining as Σ, states are defining as C, and neural network transition function defines as δ. As an example, the RNN traditional language model assumes a document as a sequence. Predicting the other word in the LSTM system takes the trained word into the previous word's account. Therefore, the LSTM framework maximizes, as shown in Equation 3.

$$q(S_t|S_{t-1}|, S_{t-2}, ..., S_0; model) \tag{3}$$

LSTM state contains input words that convert to vectors in Σ. System output demands vector and size of the words dictionary, followed by the "Softmax"žactivation function. Although, based on the size of dictionaries, challenges occur during the process.

3.3. Optimization based on Bat algorithm

Xin-She Yang first represented the bat algorithm in 2010, which is the meta-heuristic algorithm of bats based on echolocation properties. Echolocation helps in bats' flying and hunting behavior and makes them move and identify various insects in totally dark places. There are generally three rules presented by Xin-She Yang [52] while bat algorithm implementation:

- The first one is related to distance sensing, which all bats use echolocation. They can also identify similar or non-similar among with food and background in a way.
- Bats fly randomly with velocity c_i at position u_i with a fixed frequency d varying wavelength m and loudness G_0 to search for prey. one of the bat's ability is to adjust the emitted pulse wavelength and redact the pulse emission I based on target proximity to a range of [0, 1].
- As long as loudness can modify in multiple ways, we assume the loudness range from massive G_0 to minimum fixed value G_{min}.

Bat algorithm optimization presented as a pseudocode in Algorithm 1.

Algorithm 1. Pseudo Code of Bat Algorithm

1: Initialize BA and problem specific parameters
2: Objective function define as $f_u, u = (u_1, u_2, u_3, ..., u_d)^T$
3: Bat population initialize as c_i and u_i
4: Pulse frequency define as $Q_i \in [Q_{min}, Q_{max}]$
5: Pulse rate initialize as r_i and Loudness define as G_i
6: **if** $t < T_{max}$ **then**
7: Select a solution among the best solutions
8: **end if**
9: Generate new output
10: Update velocity
11: Frequency regulation to produce new solution
12: **if** r_i is < rand (0, 1) **then**
13: Recommend the solution out of outputs
14: **end if**
15: **if** $f_c < f_c$ and $G_i >$ rand (0, 1) **then**
16: Accept the solution
17: Decrease r_i and G_i
18: **end if**
19: Best present detected and bats ranked
20: End
21: Display

Each practical suppose to move with a specified velocity to attain the highest value, which returns by the objective function. Following Equation 4 evaluate and update the velocity of each iteration.

$$C_i^t = C_i^{t-1} + (C_i^{t-1} * u_*) f_i \tag{4}$$

4. Hybrid architecture based on recurrent neural network and topic modeling (LSTM-LDA)

The proposed system comprises three main modules, i.e., the machine learning process, the combination of Latent Dirichlet Allocation (LDA) topic modeling and LSTM, and finally, Knowledge discovery. The presented approach bridges the gap between LSTM and traditional latent Dirichlet allocation (LDA) topic modeling. The proposed system's primary goal is to overcome the problem statement on focused modules and required solutions. Hence, for the explained task, ideal model-quality features are needed, e.g., short group of parameters, simple interpretation, and capable of accurate prediction for future movements. Combined model defined as following Equation 5.

$$\log q_{(S)} = \log \sum_{E_1:T} \Pi_T q(S_t|E_t) q(E_t|E_{t-1}, E_{t-2}, ..., E_1) \tag{5}$$

4.1. Model structure

This section proposed the whole model structure. In presented system LSTM is applied for topic sequences in Equation 6:

$$q(E_t|E_{t-1}, E_{t-2}, ..., E_1) \tag{6}$$

and LDA is applied for word sequences in Equation 7.

$$q(S_i|E_i) \tag{7}$$

Table 2 represents used notations in proposed system and Figure 3 represents the proposed system architecture.

Table 2
Notation used in this paper

Notation	Meaning
H	Number of topics
R	Dictionary size
N	Number of Documents
L	Number of words in document
σ	Topic weight
λ	Topic probability
E	Topic assignment
c_i	Velocity
u_i	Position
d	Frequency
G_0	loudness

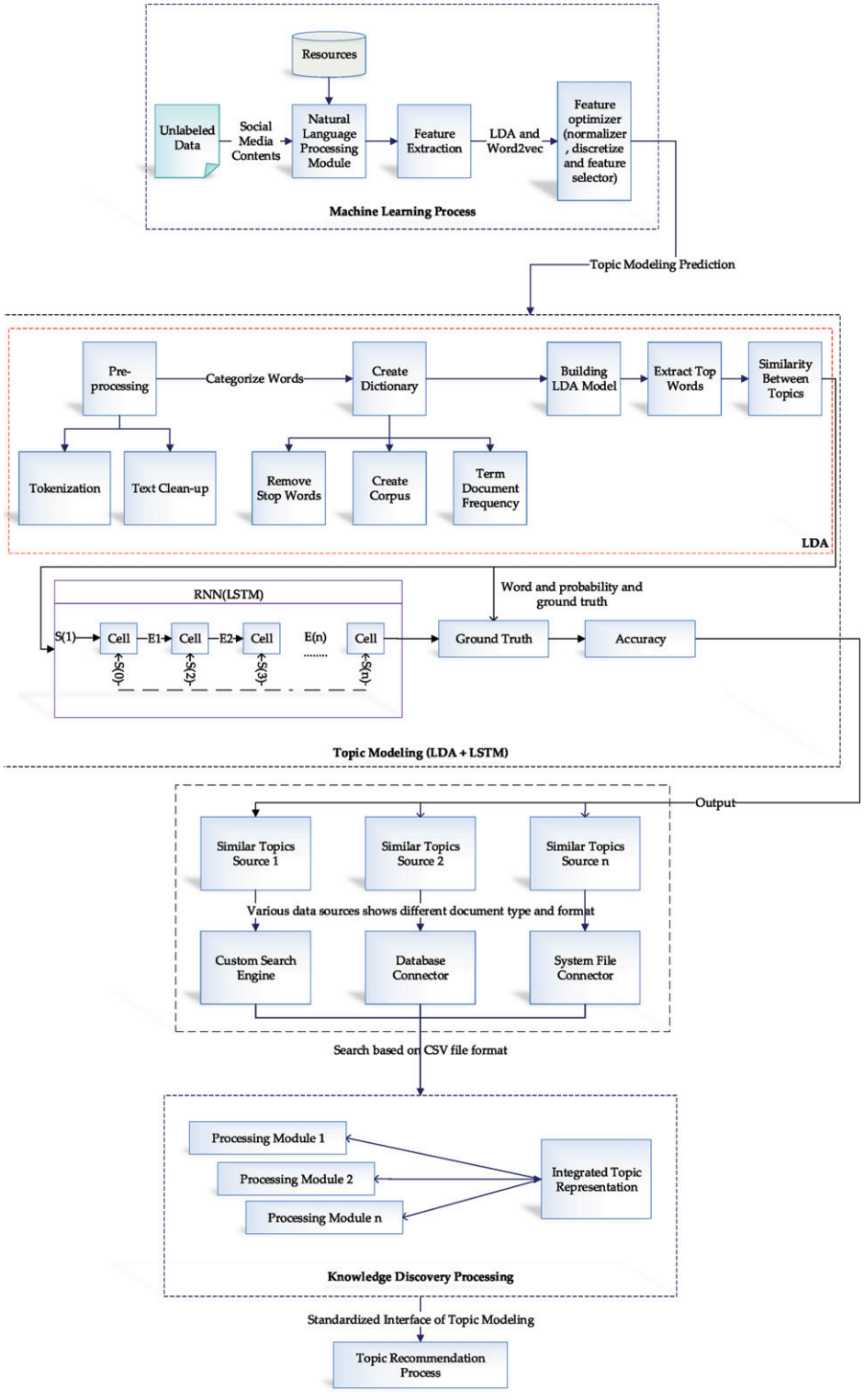

Fig. 3. System Architecture of Knowledge Discovery Process.

The proposed Figure 3 represents the input data as social media contents (tweeter, comments, news, etc.). The process starts by using machine learning techniques (XGBoost and Random Forest, etc.) to extract topics from short text datasets collected from social media websites. The topic extraction process is running based on a combination of LSTM, LDA, and word2vec modules. LDA is one of the famous areas in the topic modeling system. It is based on the word probability of occurring, and word2vec is the dictionary of words used to find words' relationships easily. At the end of the proposed system, knowledge discovery is applied to show the hidden part of the collected contents, which is the main issue in short texts.

Architecture input S_t is generated for vector in time t and LDA process shown as E_t which is latent vector. The latent vector E_t fixed in LSTM system. LSTM system evaluates the topic groups of any provided short text contents after the training process. Data ground truth is a combination of short texts besides topic labels.

LSTM based topic modeling is applied to the prediction module to predict the class of topics in the next word out of contents. After pre-processing, the input text-transform into word vectors. LDA extracts topics and, based on the pre-trained model, LSTM train the model. The next module is the knowledge discovery focused module, which is the topic recommendation that proposes hidden topics that are discovered out of texts by combining investor's priority.

4.2. Problem formulation

In this section, the problem formulation in the proposed topic prediction and knowledge discovery is evaluated. The proposed system's main problem is the lack of infomation in the short texts, making it difficult for users to understand its exact meaning. In this process, the integrated method of topic modeling, machine learning, and knowledge discovery is used to extract the hidden information of short texts and, based on topic modeling, categorize the information in proper groups. The identification rate, substitution rate and rejection rate presented in Equation 8, 9 and 10.

$$identification = \frac{correct.segments}{total.segments} * 100\% \quad (8)$$

$$subsituation = \frac{incorrect.segments}{total.segments} * 100\% \quad (9)$$

$$rejection = 100\% - identification - subsitation \quad (10)$$

To extract the useful information and hidden topics from the text, the correct segments and similarly incorrect segments require to evaluate the total number of segments. Based on this process, the number of rejected and identified segments can easily be estimated. The topic coherence, word probability, and documents similairty evaluated in Equation 11.

$$Y(t, X_t) = \sum_{n=2}^{N} \sum_{i=1}^{n-1} log(\frac{B(W_n^t, W_1^t) + 1}{B(W_1^t)}) \quad (11)$$

Here, $X_t = (W_1^t, ..., W_N^t)$ presents the list of topics with more coherent words as N. B_W shows the number of total documents with the total words W and $B(W, W')$ shows the number of which contains the co-occur words.

5. Implementation and experimental results

This section evaluates detailed information about the proposed environment and determines the dataset and experimental settings.

5.1. Experimental setting

The experimental setup of the proposed system is summarized in Table 3. The system's experiments and results are carried out using Intel(R) Core(TM) i7-8700 CPU @3.20GHz 3.19 GHz processor with 32 GB memory. The integration of LDA, LSTM, and word embedding features are used to find the relevant words and categorize them into relevant topics. Similarly, the library and framework used in the proposed system is the Jupyter notebook. The programming language used in the designing of this system is WinPython–3.6.2.

5.2. Data set

This section is to show the process of collecting data from social media content.

Table 3
Development Environment of Proposed Topic Recommendation

Component	Description
Programming language	WinPython–3.6.2
Operating system	Windows 10 64bit
Browser	Google Chrome, opera
GPU	Nvidia GForce 1080
API	Tensorflow
Library and framework	Jupyter notebook
CPU	Intel(R) Core(TM) i7-8700 CPU @3.20GHz 3.19 GHz
Memory	32GB
Recommendation Modules	LSTM and latent
Dirichlet allocation (LDA)	

5.2.1. Dataset collection

Typically, the collected dataset from different resources is incomplete. Lack of attribute value, lack of interest attributes, or data is aggregate. Based on applying a machine learning system to extract topics and knowledge, collected data are from social media contents using data miner extension. The dataset contains comments, emails, daily news broadcasts on tweeter, Facebook, etc. Dataset collection defines three main steps, which are shown in Figure 4. One of the required extensions is based on broadcast news, e.g., articles, discussions, or social media websites, e.g., tweeter, Facebook, etc. This process is the primary step to collect URLs and make a list of data. Furthermore, we supplement our collected dataset by searching for news from Google, Facebook, Reddit, etc.

5.2.2. Experimental data

Table 4 shows all detailed information related to the dataset and proposed experiments. 12,600 URLs collected from social media contents contain short texts or information related to news and any kind of comments. We filter all collected contents to not hidden documentation that is available in public. After the filtering process, the left contents are 11,314. 70% of the whole dataset used for training data and 30% for testing dataset.

Table 4
Experimental Environment Implementation

Data Characteristics	Specifications
Total No. of collected URL's before filtering	12,600
Total No. of collected URL's after filtering	11,314
Training data	70%
Test data	30%
Splitting method	Topic-based split

Table 5
LDA analysis for topic identification classes

	Data	Explanation
T0	autos	key, use, government, law, right, state
T1	hardware	max, car, gun, write, article
T2	graphics	window, use, line, drive, subject
T3	space	god, write, say, people, subject
T4	politics	say, people, armenian, know, write
T5	sci	space, people, think, say, organization
T6	windows	use, image, file, space, available
T7	motorcycles	line, subject, organization, post, host
T8	religion	organization, subject, line, write, article
T9	forsale	game, team, line, play, year

5.3. Model training process

A combination of the LSTM-LDA model extracts topics from the mentioned dataset and defines labels for extracted topics based on the information and the highest probability of words. As it is shown in Table 5 Labels in this system are named az T0, T1, T2, T3, T4, T5, T6, T7, T8, and T9. LDA is one of the prominent features use for the topic selection process, and all the topic extraction steps in this paper implement using the Gensim library in python. Next is training with LSTM, which classifies contents into topics.

Figure 5 shows the training process for topic extraction. The first part presents the LDA input analysis. Based on the topic classification, clustering, and feature extraction process, the detailed information presented in Table 5, topics extracted, and the next step labeling topics through the output information. Finally, topic weights show the probability of extracted topics comparing with available contents.

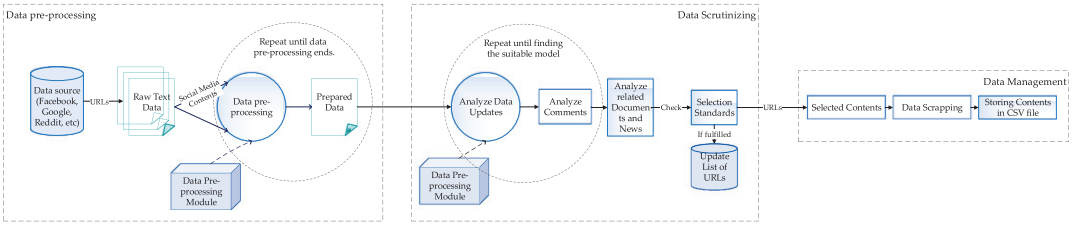

Fig. 4. Data Collection Process.

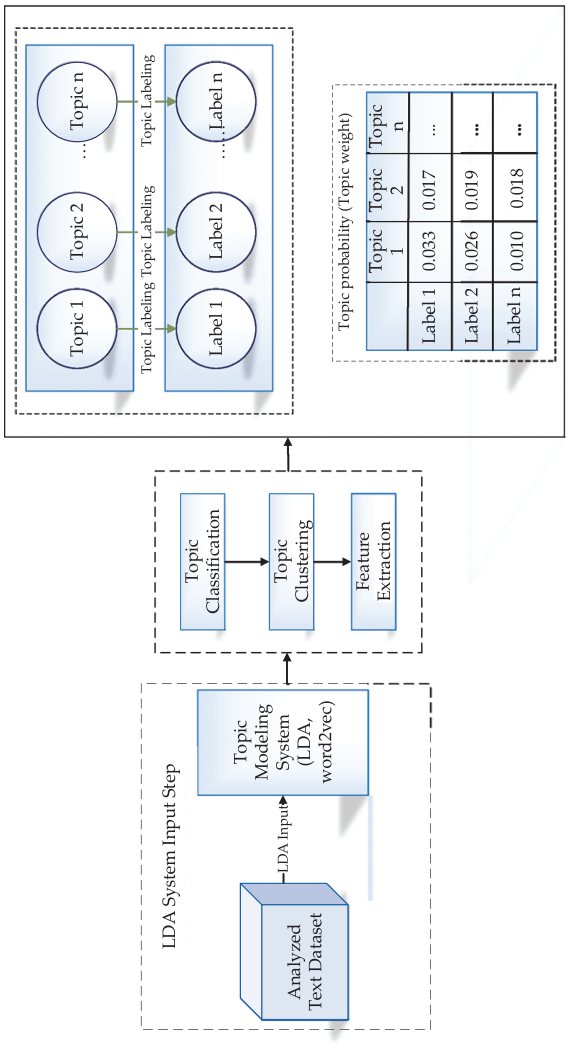

Fig. 5. Topic Modeling Training Process.

5.4. Optimized knowledge recommendation system

In a simple explanation, the optimized knowledge recommendation system is proposed to recommend the hidden part of context by operating the output result of topic modeling prediction based on the investor's priority. recommendation steps present as below:

- Higher reliability contents are selected.
- User priority checked and categorize contents.
- Recommendation system shows the new result based on extracted knowledge.

Concerning to find the highest reliability of content optimization function is needed. Figure 6 displays the knowledge recommendation process, which shows the hidden part of the contents that fit the optimization module. The used optimization module in the proposed system is the Bat algorithm. This algorithm input is listed as user priority, topic modeling prediction, knowledge recommendation, and a group of restrictions. The optimization algorithm is fundamental for the objective function to show the output. The objective function aims to recommend the hidden knowledge out of available contents based on user priority in the presented system.

5.4.1. Knowledge credibility estimation

Estimating knowledge credibility shows us how to find the highest probability of optimal knowledge recommendation. The highest probability of reliable knowledge presentation can be presented as contents that cross from the user's priority with the maximum probability of on-time delivery. This part shows the contents of dependency and credibility. Contents reliability obtain from communication patterns based on contents update, e.g., used keywords, investor's view, reward or delivery assurance, etc. Contents reliability calculated based on Equation 12. URLs related features used for contents creation and dependent elements.

$$Reliability_{content} = [\sum_{i}^{n} \frac{Type_{A_i}}{Type_{D_i}} + \frac{URLs_{social}}{R_{score} + delay_{post}}] - (M_w + N_w) \quad (12)$$

where M_w and N_w are presented as $TypeA$ weight. Table 6 shows parameter definitions.

Each URL contains various contents that categorize into several topic classes. Accordingly, the percentage of contents deliberated. For a more straightforward process, the total number of contents is divided into three key points, named Type A, Type B, and Type C. Type A contents contain overhang and critical texts. Thus, Type A contents are separated from analyzed data to control the recommendation process from unnecessary information.

5.4.2. Optimal topic recommendation based on objective function

In the proposed study, the objective function aims to search and detect topics with a higher value of user preference and high reliability, i.e., topics with high validity. Hence, the following steps show the process of projecting an objective function:

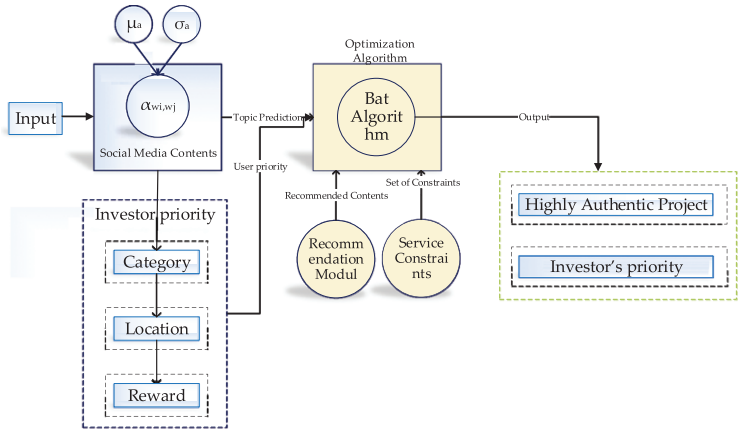

Fig. 6. Optimized knowledge recommendation module based on Bat Algorithm.

Table 6
Parameter reliability definition

Parameters Reliability		Description	Notations
		Content-based	
1	Type A	Contents weight from T3 to T5	M_w
2	Type B	Contents percentage from T0 to T2 & T6	$Type_B$
3	Type C	Contents percentage from T7 to T9	$Type_C$
4	Readability score	Content clarification measure	R_{score}

- Increase the higher reliability (i.e., the maximum weight of T7, T8, T9)
- User preference higher options
- Reduce the level of lower reliability (i.e., Reduce the weight of T0, T1, T2, T3, T4, T5, T6)

Afterwards,

$$weight_1 = \gamma(T7) + \varepsilon(T8) + \vartheta(T9) + x(Userpreference) \quad (13)$$

$$weight_2 = \theta(T0) + \theta_1(T1) + \theta_2(T2) + \theta_3(T3) + \theta_4(T4) + \theta_5(T5) + \theta_6(T6) \quad (14)$$

In Equation 13, θ, ε, ϑ, and x display the weight of T7, T8, T9 and user preference. likewise, Equation 14 presents the weight of T0, T1, T2, T3, T4, T5 and T6 as $\theta, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ and θ_6. Objective function based on Bat Algorithm is responsible to reduce the weight of topic classes from T0 to T6 and increase user preference weight from T7 to T9. The presented process evaluate as below in Equation 15:

$$W = Increase_{weight_1} + Decrease_{weight_2} \quad (15)$$

Bat Algorithm process summarized in Figure 7.

In the Bat algorithm, positions are selected randomly to evaluate the position and fitness of the algorithm. Estimated fitness help to evaluate the objective function which is Increase $weight_1$ and Decrease $weight_2$. For the next step, based on the evaluated position and fitness, the pulse rate is generated. Consequently, for the next step, current fitness is compared with the pulse rate. If the Current fitness is better than r_i, then r_i updated to selecting the best solution randomly. Else it evaluates the new fitness. Step forward, New fitness compared with loudness G_i. If the new fitness is smaller than loudness and smaller than previous fitness, it updates the pulse rate and loudness. If the new fitness is smaller than the best fitness, it updates the bat position and best fitness. Else it stays the same and terminates.

6. Results

This section presents the final result of the proposed hybrid approach regarding topic prediction in social media content. Based on these predictions, the implemented process shows the knowledge recommendation based on the topic's hidden parts. The proposed system recommends highly reliable topics for online users. In the next section, the LSTM-LDA prediction system accuracy is presented.

6.1. Prediction accuracy of optimized recommendation module

In this section, LSTM-LDA prediction is compared with other baselines mentioned as simple Neural Networks or (NNs), which shows as NN-LDA.

6.1.1. Prediction accuracy of topic classes

Figure 8 shows the comparison of prediction accuracy between different baselines. In this system, the main LSTM-LDA (RNN-LDA) model is compared with the LDA simple Neural Network Model in training set. The reason of providing training set result is because the model is non-linear and it process based on training set only. It is noticed that the accuracy of RNN-LDA is relatively better than NN. The RNN-LDA model's accuracy is almost 96% as a contrast with Neural Network (NN), which is 92%, and NN-LDA is 94%.

6.1.2. Prediction accuracy for number of topics

Figure 9 shows the LDA topic modeling training process based on prediction on topic classes. This process helps to get a higher probability of topics and the best prediction results. Defined topics are from 0 to 9. Based on Figure 9 details, topics accuracy between the three models are different for each topic. The maximum accuracy achieved in this pro-

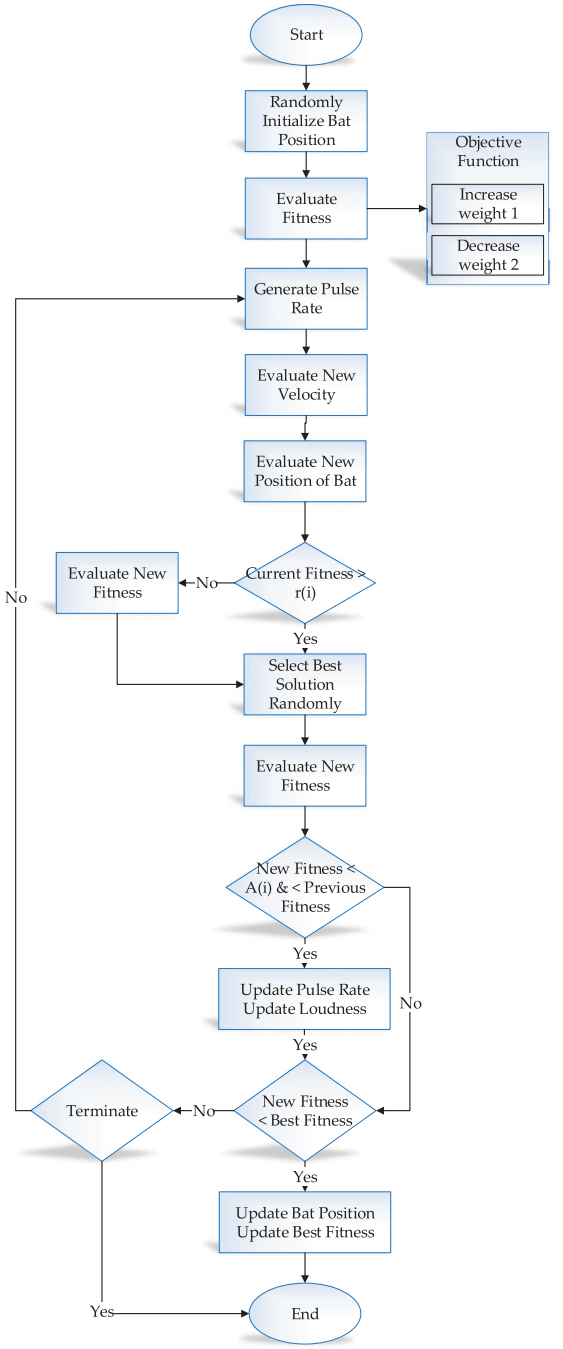

Fig. 7. Optimized recommendation based on Bat Algorithm flow chart

cess is between topics 5 to 8. All presented algorithm in topic 8 achieves to their highest prediction process accuracy.

Table 8 presents the LDA system's detailed information and the number of dominant topics from topic 1 to 9.

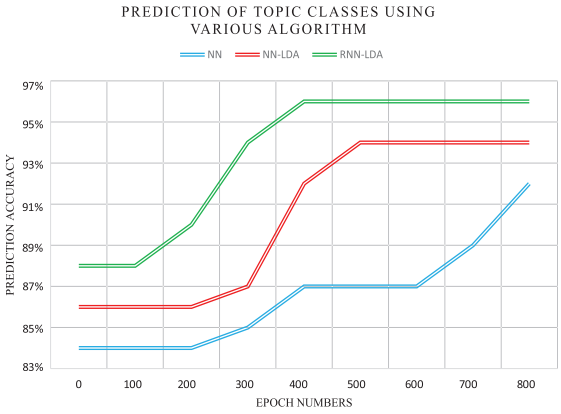

Fig. 8. Prediction Accuracy of Topic Classes Using Various Algorithms

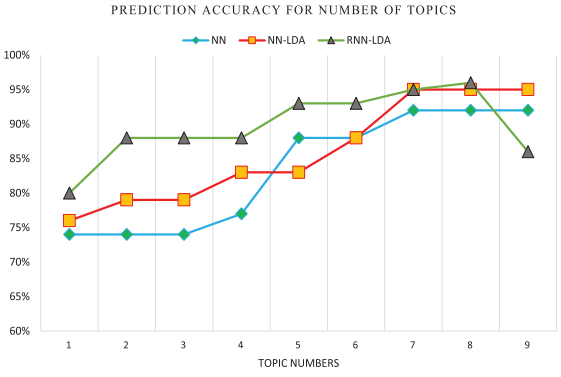

Fig. 9. Prediction Accuracy for Number of Topics

$T0$ to $T9$ are representing the number of topics. Selected parts are showing the highest probability of the document in each topic. If the document has any information, it shows as mentioned numbers. Else it is shown as zero. Dominant topics are representing the subtopics for each topic number. Simply explaining it means that each topic is also divided into subtopics that show one category is a mixture of a similar subtopics number. Table 8 shows the number of processed documents in each topic. The total number of the document collected from famous Internet websites, after processing is 11,314.

Figure 10 presents topic visualization details. Each circle is representing one topic which in the proposed approach, we defined nine topics. The distance between topics shows the similarity between topics. In this process, 30 relevant words are shown on the right side of the figure. By clicking in each circle, details of that topic are shown on the right side. The

Table 7
Dominant Topics in Each Topic Number

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Dominant T
D0	0	**0.61**	0	0	0	0	0	**0.38**	0	0	1
D1	0.03	0	**0.79**	0	0	0.09	0	0	0	0.08	**2**
D2	0	**0.15**	**0.4**	0.06	0.07	0.02	0	**0.21**	0	0.08	**2**
D3	0	0	**0.31**	0	0	0	0	**0.53**	**0.15**	0	**7**
D4	0	0	**0.18**	**0.14**	0	0	**0.28**	**0.39**	0	0	**7**
D5	**0.54**	**0.24**	0	0.02	**0.2**	0	0	0	0	0	0
D6	0	0	0	0	0	**0.24**	0	**0.5**	**0.12**	**0.13**	**7**
D7	0	0	**0.95**	0	0	0	0	0	0.05	0	**2**
D8	0	0	**0.43**	0	0	0	0	**0.54**	0	0	**7**
D9	0	0	**0.75**	0	0	0	0.07	0	0	**0.17**	**2**
...	...	...	...	...	...	...	...	...	...	...	...

Table 8
Topic Distribution Review across Documents

	Topic No.	Doc No.
0	7	2095
1	2	1764
2	3	1536
3	1	1196
4	9	1122
5	8	1011
6	0	742
7	4	651
8	6	604
9	5	593

blue color presents the relevant words in the whole dataset, and by clicking on each topic, the red color shows how many probabilities of words are on that topic.

6.1.3. Comparison of machine learning algorithms

We compare six machine learning algorithms with their score, training time, and prediction time in the proposed approach. Table 9 shows detailed information about machine learning algorithms prediction results. These techniques were processed and tested in the presented dataset of this research. The results show that this procedure presents a hybrid algorithm that has higher accuracy than other machine learning models. The KNN algorithm with the 91.9% and XGBoost with the 85.6 % is in the second and third stage results.

6.2. Limitation and shortcomings of the proposed solution

There are some limitations and problems in the proposed work mentioned as following: The first thing is the obvious drawback of topic models, which shows the probability of words in the topic based on the number of documents. If the number of documents is less, then the topic probability also assigns a small number of topics. This process is the same in the number of words in the document too, which automatically the topic detection provides insufficient information. Second is the proposed method only tested in English type of articles. Still, other researchers did topic modeling in other languages than English, e.g., Chinese, Arabic, etc. Third, this system trained and tested in the limited number of documents mentioned in dataset information, and huge dataset processing is not applicable. For processing a large number of the dataset, renormalization is required.

7. Discussion and conclusion

This section describes the challenges and goals of the proposed knowledge recommendation module in social media content. The main contribution of this paper is listed as:

- The first step needs the collected social media dataset (‘Tweets,’ ‘Comments,’ ‘News,’ etc.)
- The second is extracting useful information.
- The third is generalizing topic modeling based on word probability by combining LDA and LSTM.
- The fourth is solving sparsity using machine learning algorithms
- Finally, using knowledge discovery to extract essential and useful information from contents.

This process progressed to find the hidden part of knowledge and recommend reliable information to readers. A combination of LDA-LSTM hybrid

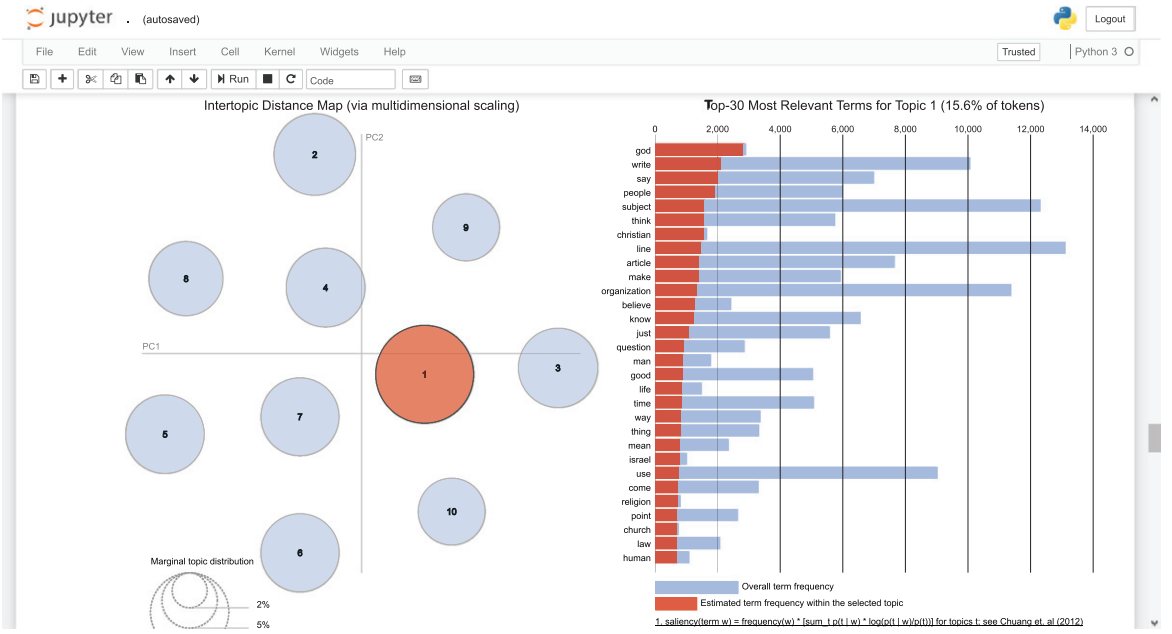

Fig. 10. pyLDAvis Topic Visualization

Table 9
Comparison of Machine Learning Algorithms

Model	Score	Training Runtime	Prediction Runtime
KNN [25]	91.9	1.01	0.12
XGBoost [53]	85.6	1.22	0.13
SVC [24]	83.6	1.26	0.08
Random Forest [54]	82.0	1.07	0.02
naive Bayes [54]	67.0	1.001	0.05
Hybrid	95.5	1.03	0.11

approach and topic modeling presented to obtain time association for topic modeling and prediction that can extract the topics out of contents. The proposed model increased the prediction system's accuracy based on the LDA model and applied a recommendation strategy based on reliable content. The idea of using a knowledge recommendation system is not a repetitive topic. Many other approaches focus on finding the relationship between topics, finding the similarity between them, or using different optimization algorithms for topic modeling. Still, the proposed approach causes an in-depth study to understand the contents and discover new information. There are some challenges in the proposed development environment. The main challenge is the ground truth data collection and the verification of it. The topic recommendation's main goal in this system is to specify the contents before recommendation to a user based on user preferences.

Acknowledgment

This research was supported by the 2021 scientific promotion program funded by Jeju National University.

References

[1] P. Montebruno, R.J. Bennett, H. Smith and C. Van Lieshout, Machine learning classification of entrepreneurs in british historical census data, *Information Processing & Management*, **57**(3) (2020), 102210.
[2] B. Marr, A short history of machine learning–every manager should read, *Forbes* http://tinyurl. com/gslvr6k, 2016.
[3] Y. Hong, B. Hou, H. Jiang and J. Zhang, Machine learning and artificial neural network accelerated computational discoveries in materials science, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **10**(3) (2020), e1450.

[4] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M.M. Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine, *Journal of The Royal Society Interface* **15**(141) (2018), 20170387.

[5] J.N. Kutz, Deep learning in fluid dynamics, *Journal of Fluid Mechanics* **814** (2017), 1–4.

[6] B. Paria, C.-K. Yeh, I.E.H. Yen, N. Xu, P. Ravikumar and B. Póczos, Minimizing flops to learn efficient sparse representations, *arXiv preprint arXiv:2004.05665*, 2020.

[7] P. Xie, D. Yang and E. Xing, Incorporating word correlation knowledge into topic modeling, In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, pages 725–734, 2015.

[8] A. Linsel, K. Bär, J. Haas, J. Hornung, M.D. Greb and M. Hinderer, Georevi: A knowledge discovery and data management tool for subsurface characterization, *SoftwareX* **12** (2020), 100597.

[9] W.J. Raynor, Knowledge and data discovery management systems (kddms), In *The International Dictionary of Artificial Intelligence*, pages 154–154. Routledge, 2020.

[10] C. Ferner, C. Havas, E. Birnbacher, S. Wegenkittl and B. Resch, Automated seeded latent dirichlet allocation for social media based event detection and mapping, *Information* **11**(8) (2020), 376.

[11] K. Kumar, Probabilistic latent semantic analysis of composite excitation-emission matrix fluorescence spectra of multicomponent system, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, page 118518, 2020.

[12] M. Hoffman, F.R. Bach and D.M. Blei, Online learning for latent dirichlet allocation, In *advances in neural information processing systems*, pages 856–864, 2010.

[13] P. Xie and E.P. Xing, Integrating document clustering and topic modeling, *arXiv preprint arXiv:1309.6874*, 2013.

[14] J. Wan, J. Li, Q. Hua, A. Celesti and Z. Wang, Intelligent equipment design assisted by cognitive internet of things and industrial big data, *Neural Computing and Applications* **32**(9) (2020), 4463–4472.

[15] H.M. Pandey, N. Bessis, S. Das, D. Windridge and A. Chaudhary, Editorial to special issue on hybrid artificial intelligence and machine learning technologies in intelligent systems, 2020.

[16] B. Molina-Coronado, U. Mori, A. Mendiburu and J. Miguel-Alonso, Survey of network intrusion detection methods from the perspective of the knowledge discovery in databases process, *arXiv preprint arXiv:2001.09697*, 2020.

[17] M.E. Günay and R. Yıldırım, Recent advances in knowledge discovery for heterogeneous catalysis using machine learning, *Catalysis Reviews*, pages 1–45, 2020.

[18] C. Sievert and K. Shirley, Ldavis: A method for visualizing and interpreting topics, In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[19] J. Murdock and C. Allen, Visualization techniques for topic model checking, In *AAAI*, pages 4284–4285, 2015.

[20] D.M. Blei, Probabilistic topic models, *Communications of the ACM* **55**(4) (2012), 77–84.

[21] Marc'Aurelio Ranzato and M. Szummer, Semi-supervised learning of compact document representations with deep networks, In *Proceedings of the 25th international conference on Machine learning*, pages 792–799, 2008.

[22] H. Larochelle and S. Lauly, A neural autoregressive topic model, *Advances in Neural Information Processing Systems* **25** (2012), 2708–2716.

[23] A. Behera, *Combination of topic modelling and deep learning techniques for disaster trends prediction*. PhD thesis, Dublin, National College of Ireland, 2019.

[24] R. Albalawi, T.H. Yeap and M. Benyoucef, Using topic modeling methods for short-text data: A comparative analysis. front, *Artif, Intell* **3** (2020), 42.

[25] A. Bagheri, A. Sammani, P. Van D. Heijden, F.W. Asselbergs and D.L. Oberski, Etm: Enrichment by topic modeling for automated clinical sentence classification to detect patients' disease history, *Journal of Intelligent Information Systems*, 2020.

[26] T. Baltrušaitis, C. Ahuja and L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2) (2018), 423–443.

[27] N.A. Huhnstock, A. Karlsson, M. Riveiro and H.J. Steinhauer, An infinite replicated softmax model for topic modeling, In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 307–318. Springer, 2019.

[28] J.C. Crotts, P.R. Mason and B. Davis, Measuring guest satisfaction and competitive position in the hospitality and tourism industry: An application of stance-shift analysis to travel blog narratives, *Journal of Travel Research* **48**(2) (2009), 139–151.

[29] E. Elmurngi and A. Gherbi, Detecting fake reviews through sentiment analysis using machine learning techniques, *IARIA/data analytics*, pages 65–72, 2017.

[30] A. Shukla, W. Wang, G.G. Gao and R. Agarwal, Catch me if you can—detecting fraudulent online reviews of doctors using deep learning, *Ritu, Catch Me If You Can—Detecting Fraudulent Online Reviews of Doctors Using Deep Learning (January 14, 2019)*, 2019.

[31] Z. Xiang, Q. Du, Y. Ma and W. Fan, A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism, *Tourism Management* **58** (2017), 51–65.

[32] L. Chen, W. Li, H. Chen and S. Geng, Detection of fake reviews: Analysis of sellers' manipulation behavior, *Sustainability* **11**(17) (2019), 4802.

[33] X. Wang, L.R. Tang and E. Kim, More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? *International Journal of Hospitality Management* **77** (2019), 438–447.

[34] Y. Chaudhary, P. Gupta and T. Runkler, Lifelong neural topic learning in contextualized autoregressive topic models of language via informative transfers, *arXiv preprint arXiv:1909.13315*, 2019.

[35] Y. Jo, L. Lee and S. Palaskar, Combining lstm and latent topic modeling for mortality prediction, *arXiv preprint arXiv:1709.02842*, 2017.

[36] W. Shafqat and Y.-C. Byun, Topic predictions and optimized recommendation mechanism based on integrated topic modeling and deep neural networks in crowdfunding platforms, *Applied Sciences* **9**(24) (2019), 5496.

[37] P. Jansson and S. Liu, Topic modelling enriched lstm models for the detection of novel and emerging named entities from social media, In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4329–4336. IEEE, 2017.

[38] N. Kawamae, Topic structure-aware neural language model: Unified language model that maintains word and topic

ordering by their embedded representations, In *The World Wide Web Conference*, pages 2900–2906. ACM, 2019.

[39] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[40] C. Lai, M. Farrús and J.D. Moore, Integrating lexical and prosodic features for automatic paragraph segmentation, *Speech Communication*, 2020.

[41] D. Griol, J.M. Molina, A. Sanchis and Z. Callejas, A data-driven approach to spoken dialog segmentation, *Neurocomputing* **391** (2020), 292–304.

[42] Z. Shahbazi and Y.-C. Byun, Analysis of domain-independent unsupervised text segmentation using lda topic modeling over social media contents, *International Journal of Advanced Science and Technology* **29** (2020), 5993–6014.

[43] Z. Shahbazi, F. Jamil and Y. Byun, Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning, *Journal of Intelligent & Fuzzy Systems*, (Preprint) (2020), 1–18.

[44] F.M. Bianchi, D. Grattarola and C. Alippi, Spectral clustering with graph neural networks for graph pooling, In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020.

[45] X. Li, C. Li, J. Chi and J. Ouyang, Short text topic modeling by exploring original documents, *Knowledge and Information Systems* **56**(2) (2018), 443–462.

[46] Z. Shahbazi, Y.-C. Byun and D.C. Lee, Toward representing automatic knowledge discovery from social media contents based on document classification, *Int J Adv Sci Technol*, 2020.

[47] Z. Shahbazi and Y.C. Byun, Toward social media content recommendation integrated with data science and machine learning approach for e-learners, *Symmetry* **12**(11) (2020), 1798.

[48] Z. Shahbazi, D. Hazra, S. Park and Y.C. Byun, Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches, *Symmetry* **12**(9) (2020), 1566.

[49] Z. Shahbazi and Y.-C. Byun, Product recommendation based on content-based filtering using xgboost classifier, *Int J Adv Sci Technol* **29** (2019), 6979–6988.

[50] X. Shu, *Knowledge Discovery in the Social Sciences: A Data Mining Approach*, Univ of California Press, 2020.

[51] J. Zenkert, A. Klahold and M. Fathi, Knowledge discovery in multidimensional knowledge representation framework, *Iran Journal of Computer Science* **1**(4) (2018), 199–216.

[52] X.-S. Yang, A new metaheuristic bat-inspired algorithm, In *Nature inspired cooperative strategies for optimization (NICSO 2010)*, pages 65–74. Springer, 2010.

[53] M. Khalifa and N. Hussein, Ensemble learning for irony detection in arabic tweets, In *FIRE (Working Notes)*, pages 433–438, 2019.

[54] J. Rashid, S.M.A. Shah and A. Irtaza, Fuzzy topic modeling approach for text mining over short text, *Information Processing & Management* **56**(6) (2019), 102060.

[55] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K.H. Kim and K.-S. Kwak, Transportation sentiment analysis using word embedding and ontology-based topic modeling, *Knowledge-Based Systems* **174** (2019), 27–42.