Received: November 15, 2021. Revised: February 5, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Graph Model and Deep Learning for Topic Labels in Classifying Short Texts of Scientific Article Titles

Surya Sumpeno[1,2] Diana Purwitasari[2,3]* Bastian Farandy[3] Dini Adni Navastara[3] Mauridhi Hery Purnomo[1,2]

[1]*Computer Engineering Department, Institut Teknologi Sepuluh Nopember, Indonesia*
[2]*University Center of Excellence on Artificial Intelligence for Healthcare and Society, Indonesia*
[3]*Informatics Engineering Department, Institut Teknologi Sepuluh Nopember, Indonesia*
* Corresponding author's email: surya@its.ac.id

Abstract: In an article searching system, topic categorization could guide researchers in finding the appropriate documents among the abundant availability of scientific articles. Because they are easily obtained from the internet, there is a preference for using short texts to full-text articles for data collection of the searching system. However, topic label scarcity becomes a problem, especially when preparing the system in a cold-start situation or without predefined topic categories. Typical topic analysis with a statistical-based unsupervised Latent Dirichlet Allocation (LDA) identifies clusters of words or topics based on the word distribution to overcome the label scarcity. For ease of use, words in LDA topics are manually observed, and some are set as topic names. Lower precision happened when tagging other articles using the LDA topics for the searching system preparation with categorization or classification approach. The precision values could be influenced by too many identified LDA topics. Thus, the overlapped context in the LDA results is possible since the same words appear in different topics, leading to many false positives. Here, the instigated problem is making classification results have comparable accuracy and precision values with the existing data condition of no topic labels and overlapped context when identifying topics. The problem solution is motivated to consider the word relations with others when identifying topics to differentiate the word context. Therefore, our contribution is to investigate LDA and relationships between words as a graph with a prevalent neural network model of deep learning called Graph Convolutional Network (GCN) for automatically determining the topics before examining them in classification tasks. Guided by the proposed framework, we synthesize training samples to make the dataset for LDA topics more similar in contexts. The empirical analysis through the experiments has thoroughly evaluated the LDA topics as a baseline to compare the results of statistical based (LDA) and Deep Learning based topic identification (Deep LDA) to ensure the topic quality. Then, we compared the usage of GCN with other frequently used text classifications. The classification results showed that the graph representation used to capture the subject context relationship between keywords in GCN has supported the balance between accuracy and precision.

Keywords: Topic modeling, Latent Dirichlet allocation, Graph convolutional network, Short text classification.

1. Introduction

With the increasing number of research works leading to the excess number of scientific articles available on the internet, finding the one just what a researcher needs is a tricky task. Article classification based on subject topics could assist the seeking users when utilizing an article searching system. When there are no predefined topic subjects which often occur in a cold-start situation while preparing the system, topic identification is required. Term or keyword occurrences, known as term frequency (TF), and the probability values of those words appearing in texts obtained with a statistical model of unsupervised approach Latent Dirichlet allocation (LDA) determine topics or clusters of words [1, 2]. LDA inspects the hidden thematic structure of texts. Those studies investigated topic identification from focused subjects like web documents of a selected issue crawled during three-year periods [1] or abstract texts from particular journals around two

decades [2]. Aside from those studies that consider subject homogeneity to ensure the topic quality, other studies from more diverse news article subjects but a limited number of articles have identified topics then employed them in a classification task [3]. It has LDA modifications for categorizing or classifying texts with topics as their class labels, such as adding a document ranking approach. Those works showed that LDA usage alone was not enough such that the document importance in terms of ranks is employed to filter significant words before topic identification. Other works from some domain-specific journals but still covered various research subjects investigated full-texts of scientific articles for understanding LDA topic modelling [4, 5]. Those works showed that LDA without modification gave enough coherence of topic results when extra texts were involved, such as full-text articles.

However, when preparing an article searching system in a cold-start situation, short texts of article titles are easier to collect from the internet for further studies. Thus LDA alone is not enough for handling short texts. Classification with short texts usually takes deep learning to address the subject context's insufficiency [6, 7]. Since graph as a data structure is beneficial in highlighting the subject context, the deep learning approach evolves into a neural network with spatial moves on graph nodes called graph convolutional network (GCN). The advantage of GCN is that it stores information from corresponding data, especially relationships between words. Several studies have used GCN for text classification [8], even classifying short-texts of search keywords from customers with product results [9]. However, those studies observed the classification task with standard datasets and their available class labels as topic categories. The problem in this paper has no standard dataset with topic categories as ground truth because the collected titles have no topic labels, which leads to the LDA usage before the classification task.

Article titles as short texts have a specific word limit but contain topical information. Processing article titles as a feature matrix cause contextual data problems of sparsity and ambiguity. For context clarification, some studies inserted the context layer into the deep learning architecture used in the classification task by utilizing word concepts from the knowledge base [10]. Another work has utilized the concepts from well-known association for computing machinery (ACM) library data and combined deep learning as well as machine learning for recommendation tasks based on article titles [11]. However, those studies worked on English texts. Thus, an existing knowledge base that provides pairs of words and concepts in the English language or ACM computing classification system could serve the purposes for those studies but not in our case. As a result, our proposed framework offers procedures to process non-English texts for preparing a searching system in a cold-start situation. Different studies on a recommendation for systematic reviews had a natural language approach with the recent famed deep learning architecture of bidirectional encoder representations from transformers (BERT) [12]. The studies showed that the ensemble learning model increases sensitivity and specificity values. Those studies make use of the advantage of deep learning and lean on the data abundance that our case also has. However, they relied on additional concepts, which are often unavailable in a cold-start situation.

The problem in this work originates from preparation of a searching system for scientific articles by collecting article titles as data sources from the internet, which often do not come with predefined labels of topic categories. The system is expected to help students for their final projects or researchers to study literature that includes topics extended for different domains. Similar academic search systems such as Google scholar may not be enough since there is also a need to match the topics with research areas prioritized by one nation or a university. The need to establish such system [13] with any non-English texts is often encountered. Thus, topic identification is essential before classifying or categorizing more of the title texts collected sooner or later. Previous works [1-3] demonstrated LDA for topic identification on limited domains without further classification. In contrast, other works with the topic range extended for different domains required more than short texts to ensure the topic quality [4, 5]. Hence, the deep learning approach is convenient for classifying short texts [6-9]. However, the possibility of overlapped context needs additional information from word concepts of knowledge base [10] or predefined categories of classification system [11, 12].

Our proposed framework has contributed to the challenges of non-English short-text classification tasks in a cold-start situation with subject diversities of article titles. The topic labels are defined with a generative statistical model of LDA. The proposed framework recommends procedures to determine the topics as a classification problem with approaches to justify label correctness, as well as synthesizing sampling data and supporting subject homogeneity to ensure the topic quality. The abundance of article titles compensates for the topic label scarcity. Since the abundance aspect leads to overlapped context when identifying topics, the proposed framework also considers the word relations with others as

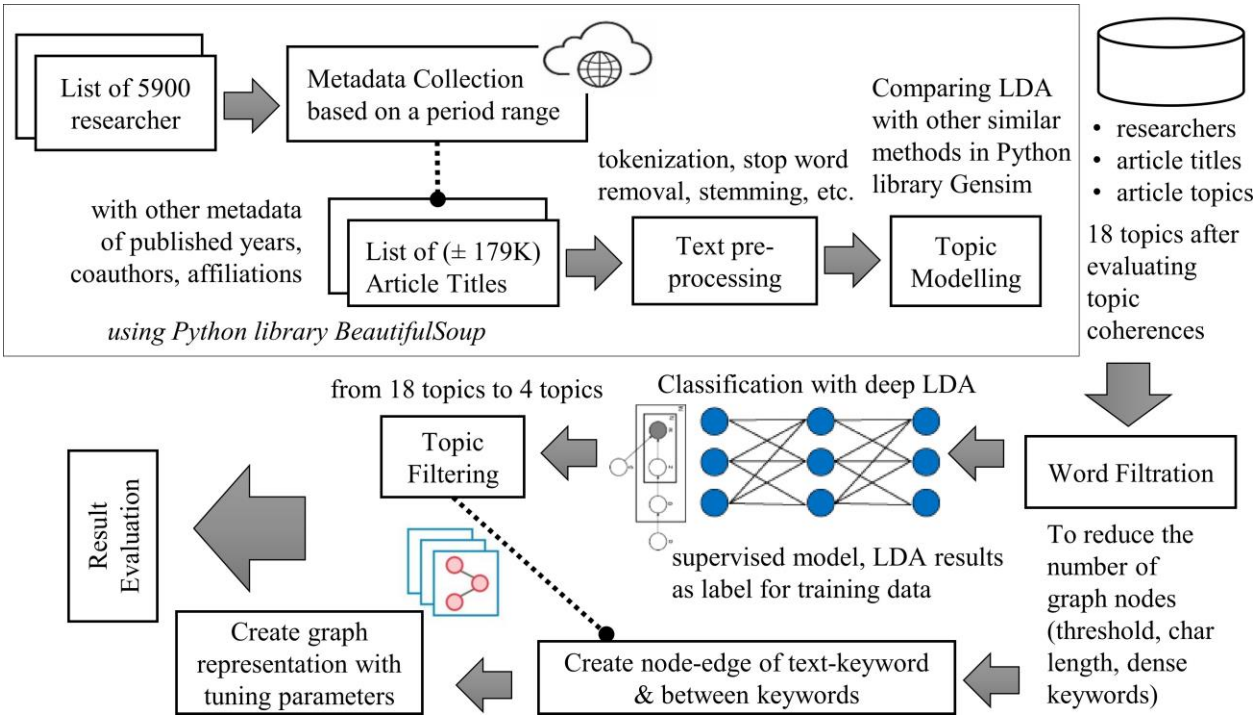

Figure. 1 Proposed framework for classifying article titles using graph model and deep learning for topic labels

graphs in the deep learning architecture (GCN).

This section (*1. Introduction*) has discussed the classification problem on short texts with abundance numbers and no prior class labels that lead to the overlapped context and could influence precisions. The discussed references have shown that our framework is about topic identification and keeping the overlapping context away on non-English short texts for classification. The following section describes the framework (*2. Proposed Method*) including detailed procedures to process the collected data (*2.1 Data Collection and Pre-processing*) and then topic identification that considering word relations before classification task (*2.2 Topic Labelling*). Afterward, several assessments for empirical analysis on mixed language short texts of Bahasa (Indonesian language) and English are described. The proposed framework processes those keywords based on their occurrence such that any translation is not required. The conducted experiments (*3. Results and Discussion*) are about graph models (*3.1*) to observe the topic quality and deep learning for classification task (*3.2*) to analyse accuracies and precisions.

2. Proposed method

Fig. 1 shows steps in our proposed framework. It starts from data collection, pre-processing until evaluation the labelling results with information about Python libraries for implementations.

2.1 Data collection and pre-processing

This work could be applied in a cold-start situation when a research management unit needs to record researchers and their expertise as assets. For avoiding the hassle of requesting lists of published scientific articles from each researcher, our framework solution starts with data collection, i.e., scraping publicly available article metadata from Google scholar of selected researchers. This work investigated the researchers who tend to be actively involved in research activities. Those researchers are mostly inclined to publish scientific articles and apply for national research grants. The selected researchers would have an affiliation to top state universities, or in our case, the ones who have legal entity status from the Indonesian ministry of research, technology and higher education.

Our dataset has 3,900 researchers included in the receiver lists of national research grants between 2018 and 2020. Researcher records could be catalogued at the national level, i.e., the Slovenian scientific system [13] who makes top researchers known to promote a productive research environment. Thus, there are an additional 500 researchers from the Indonesian scientific system called the science and technology index (SINTA) (http://sinta.ristekbrin.go.id). We scraped article metadata from Google scholar to obtain ± 209K (209,000) research articles published from 2010 to

Table 1. Data description on LDA identified topics

Details on LDA identified topics				sample keywords
Number of article titles and its percentage in a data collection with topic labels on period 2010-2015		Number of article titles and its percentage in a data collection with topic labels on period 2016-2020		
24,584 (35.7%)	organic chemistry (5,729), chemistry (4,488), agriculture (3,999), microbiology (3,899), livestock (3,239), nutrition-biotechnology (3,230)	43,082 (31.2%)	organic chemistry (6,821), chemistry (10,318), agriculture (4,868), microbiology (6,684), livestock (6,541), nutrition-biotechnology (7,850)	synthesis, acid, carbon, material
3,310 (4.8%)	ecosystem/ conservation (3,310)	8,756 (6.3%)	ecosystem/ conservation (8,756)	characteristic, area
6,405 (9.3%)	public health (3,404), disease (3,001)	14,334 (10.4%)	public health (5,780), disease (8,554)	patient, risk, factor, child
4,283 (6.2%)	environment (4,283)	5,517 (4.0%)	environment (5,517)	province, society
10,305 (15.0%)	computer eng. (3,756), mechanical eng. (3,670), electrical engineering (2,879)	22,886 (16.6%)	computer eng. (9,196), mechanical eng. (7,398), electrical eng. (6,292)	control, optimization
4,921 (7.1%)	law (4,921)	7,219 (5.2%)	law (7,219)	case, regency
3,859 (5.6%)	economic community empowerment (3,859)	5,918 (4.3%)	economic community empowerment (5,918)	influence, management
11,204 (16.3%)	education (5,112), natural resources (3371), science education (2,721)	30,470 (22.1%)	education (8,625), natural resources (10,294), science education (11,551)	student, improve, review

2020 by those researchers. The article titles are in English or Indonesian words since researchers could publish in national journals or international conferences. Thus, the short texts were identified with LangDetect library before text pre-processing.

Several typical text pre-processing steps were applied on the short texts of those collected article texts, like alphanumeric transformation, lower case changes before tokenization and stop word removal. stemming (Sastrawi library) and lemmatization (NLTK library) are used to transform words into their basic form to reduce the size of vocabulary space. Since the texts have no topic labels, some topic modelling methods [14] were observed, such as LDA or its adjustment LDA mallet, latent semantic analysis (LSA) and hierarchical Dirichlet process (HDP) that infers the topic number from the data.

Topic modelling tests with biterm topic model (BTM) regarding two words that occurred within the same context or window were also examined. However, the memory condition makes BTM is not preferable.

Various term weighting schemes to measure the term importance as keywords like the typical term frequency–inverse document frequency (TF-IDF) or LogEntropy model for normalizing the gap of TF values were also observed. We also investigated the common embedding usage of GloVe to ensure a vector representation of a text is similarly distributed in an n-dimensional space. We examined the results based on topic coherence as a performance indicator for all combinations of topic modelling with the topic number ranging from 10…90 and term weighting schemes. After comparing the topic coherence versus the processing time of those steps, our examination recommended LDA and TF-IDF with 18 as the topic number and 0.6 as the topic coherence value.

LDA could give "topic membership" or the probability value for a text to have the same topic context, such that LDA allows multi labels for the text. Since article titles have a limited number of words, usually around 15-20 words, and become less after stop word removal, it would not be easy to associate a text to a topic. Thus as parts of word filtration to reduce the vocabulary space and filter not too well-represented article titles from the dataset, our framework suggested a single topic label from the highest probability value of LDA result with a minimum threshold of 0.063. The word filtration decreases the training data from ±180K to ±109K titles. The 18 topic labels, as mentioned in Table 1, are manually defined after observing some frequent words. In this manuscript, the sample keywords have been translated into English terms.

Some topics could have related contexts, such that Table 1 shows eight topic groups. The numbers

Table 2. Average accuracy values for article title classification with Deep LDA

Notes: DO means DropOut rate	***3NN Deep LDA***			***2NN Deep LDA***		
	3 Dense DO 0.5	1 LSTM 2 Dense DO 0.5	1 LSTM 2 Dense DO 0.8	2 Dense DO 0.5	**1 LSTM 1 Dense DO 0.5**	1 LSTM 1 Dense DO 0.8
organic chemistry, chemistry, agriculture, microbiology, livestock, nutrition-biotechnology	0.62	0.64	0.63	0.61	0.65	0.64
ecosystem/ conservation	0.69	0.67	0.68	0.66	0.72	0.67
public health, disease	0.67	0.71	0.68	0.67	0.69	0.68
environment	0.53	0.60	0.51	0.48	0.61	0.52
computer eng., mechanical eng., electrical engineering	0.50	0.51	0.49	0.51	0.54	0.51
law	***0.77***	***0.77***	***0.80***	***0.77***	***0.81***	***0.80***
economic community empowerment	0.61	0.62	0.58	0.59	0.62	0.58
education, natural resources, science education	0.68	0.71	0.70	0.68	0.73	0.71

of article titles between those two periods demonstrate increasing works of researchers. Thus, a system to manage researchers' data could help profile them and be used for recommendations or visualizations, which would be a valuable tool for an institution to manage, promote and support more funding on specific research topics.

Selecting keywords from texts is an important task to analyse the context for understanding their contents. Word embedding of GloVe was also observed in our empirical analysis. The embedding usage gave better results, as described in detail in the results and discussion.

2.2 Topic labelling

We preliminary analysed the resulting LDA topic labels approximated from a dataset of ±109K titles dataset. We applied them in a frequently used deep learning-based classifier for texts: Long short-term memory (LSTM). The models were used to predict the rest of ± 29K testing data. However, the average accuracies for 18 models were relatively low even after well-established tuning parameters such as the dimension of input data set to 4,000, learning and dropout rate, or layer numbers. The proposed framework recommends two approaches for determining the correctness of topic labels in a cold-start situation based on deep learning: deep LDA [15] and GCN [8].

Experiments to determine the correctness of topic labels using deep LDA with testing data are shown in Table 2. The performance indicator is accuracy values averaged for topics with similar contexts, such as computer, mechanical, and electrical engineering. Table 2 displayed that an additional dense layer after LSTM in Deep LDA has provided better accuracy than a standalone LSTM since it helps classify texts based on output from LSTM. However, denser layers do not necessarily mean higher accuracy like 2NN vs. 3NN deep LDA showed better performance for the architecture of 1 LSTM, 1 dense, and DO 0.5. In short, 2NN deep LDA has longer training time and higher accuracy, while 3NN deep LDA has faster training time and lower accuracy.

Some categories could have widespread context than the others, like "environment" and "computer + mechanical + electrical engineering". Keywords in those categories might be commonly found in others since their subjects could be interrelated. However, the topic of "computer engineering" had better values compared to the other two topics. Their combination makes the accuracies of both categories are lower and the more specific subject like "law" provides higher accuracy. It could be inferred that some topics have better coherence than others. Thus, for determining the correctness with GCN, we focused on texts from the four selected topics with better coherence: "law, education, computer engineering, and conservation". There are ± 27,000 article titles for those topics.

The GCN usage requires a feature matrix of graph representation prepared from the dataset, $G = (V_a, V_k, E_a, E_k)$. Word filtration disregards keywords based on word length limitation and lower document frequency values to control the graph size. A keyword should have more than four letters. In a graph representation, nodes are article texts V_a and

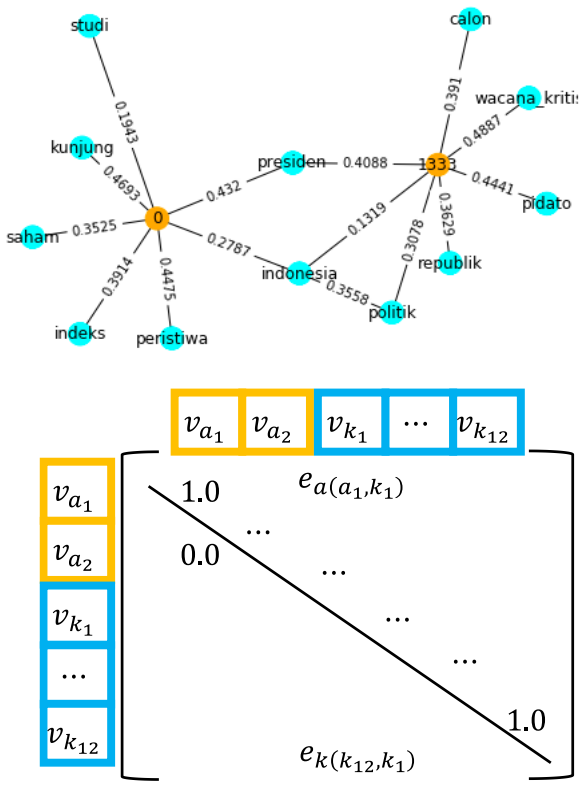

Figure. 2 A graph with texts and keywords as nodes and edges with its matrix representation

keywords V_k, including two types of edges: between keywords E_k and between an article text to keywords E_a.

With word filtration, the graph creation is only focused on keywords that occurred in the data of ±27K texts between 50…1,050 as document frequency. It could be indicated that the recommended keywords would have document frequency (DF) of 0.002-0.04 compared to the data size. Thus, the first iteration is to create edges of an article text to keywords (Eq. (1)). The edge value of E_a is TF-IDF. Then, the second iteration checks the possibility of edges between two keywords E_k by calculating their pointwise mutual information (PMI) to indicate how often those two keywords occurred in the data. Assume there are two keywords k_i, k_j, then $e_{k(k_i,k_j)}$ is computed by Eq. (2). function $TF.IDF(.)$ calculates number of term k_j occurred in text a_i combined with a log version of inverse document frequency. Function $df(.)$ is associated with the number of article titles containing certain keyword(s) or DF.

$$e_{a(a_i,k_j)} = TF.IDF(a_i, k_j) \quad (1)$$

$$e_{k(k_i,k_j)} = \log \frac{df(k_i,k_j)}{df(k_i)df(k_j)} \quad (2)$$

With an example of two article texts a_1, a_2, "Indonesian index visits the president of the study stock" as a_1 called as node-0 and "Indonesian candidates for political speech for the president of the republican critical discourse" as a_2 called as node-1333 (in Fig. 2 with normalized edge values). There are two keywords occurred in two texts: "presiden" and "indonesia" (in Indonesian language). A square feature matrix of the graph has the first rows representing article title nodes and the subsequent rows representing keyword nodes.

The feature matrix formation also has two iteration processes, first is to set the cell values as 1.0, assuming that each node is connected to itself. The second iteration is to update any cell values with the representative edge values. For example, based on Fig. 2, the size of the matrix representation is 14 x 14 with the graph density of 0.165 or the matrix sparsity of 0.847. The result of translating nodes and edges into a feature matrix for the mentioned example is displayed in Fig. 2. Then, the feature matrix becomes the input for GCN, with the output layer having the same node numbers as the number of targeted topics.

Fig. 3 demonstrates a pseudocode to clarify the classification task. Some steps with the results in Table 1 and Table 2 for justifying label correctness, synthesizing sampling data and supporting subject homogeneity to ensure the topic quality in Fig.1 have previously explained. Thus, those steps are not listed in the pseudocode (Fig. 3) since they are not parts of classification task.

3. Results and discussion

Our experiments were performed in a graphics processing unit (GPU) based cloud instance with the specifications of 8 Core CPU, 30 GB random access memory (RAM), and 8 GB NVIDIA quadro M4000. All experiments applied topics labels from LDA process on the data of ±109K short texts for 18 topics as the ground-truth dataset.

3.1 Experiments on parameters of GCN model

The experiments followed best practices [15] in setting GCN model with two convolution layers plus the same hidden node numbers (330–130), and Adam optimizer. As mentioned previously, the experiments used 4 (four) selected topics and ± 27K article titles. There are 3 (three) observations related to parameters for obtaining better GCN models.

INPUT:
a. List of (manually selected) productive researchers who received national research grants
b. List of unlabelled article titles with at least one author exists within list of researchers. The titles are scraped from Google Scholar pages of researchers. The dataset is divided into data training (± 180K) and data testing (± 29K).

OUTPUT:
a. List of topics (or clusters of words) with topic name is set to auto-id (i.e. topic-x) with better coherence values
b. List of article titles from data testing which are already labelled with topic name

PROCEDURE:
Step #01: Pre-process (tokenizing, stop-word removing, stemming) title texts and compute term weights
Step #02: FOR topic number (tn) = 10… 90 DO
Step #03: Model tn topics using LDA on the data training
Step #04: Get selected title texts with the probability to its mapped topic(s) > 0.063
Step #05: Get a topic number with coherence values > 0.6, list of topic names, their significant terms in topics and selected title texts
(*The results: 18 topics and ±109K titles from data training*)
Step #06: Get terms or keywords with Document Frequency = 0.002-0.04
Step #07: Create a graph with title texts and keywords based on a configuration in Fig. 2
Step #08: FOR each cell in the graph DO
Step #09: Calculate $e_{a(a_i,k_j)}$ or $e_{k(k_i,k_j)}$ based on Eq. (1) or Eq. (2) accordingly
Step #10: Set the cell value = 0 based on thresholds of Eq. (3) or Eq. (4) accordingly
Step #11: Build a classification model with GCN using the created graph in Step #07 and the topic labels in Step #05
Step #12: Use the classification model in Step #11 to predict the topic labels of data testing

Figure. 3 Pseudocode for classification task in the proposed framework

Table 3. Average accuracy (in %) values for article title classification with edge thresholds

scen.	$thres_{TF.IDF}$	$thres_{PMI}$	# Edges	Accuracy
1		0.2	258,429	88.77
2		0.4	257,561	88.60
3	0.2		246,867	87.97
4	0.4		129,613	70.05
5	0.2	0.2	246,143	**88.06**

a. Observing Learning Rate and Gamma Value

Learning rate controls how much the model changes in estimating error values when the node- weights updating happens, while Gamma value is a constant for learning rate multiplication to accelerate the model stabilization. The values of learning rate ranges from 0.10 – 0.05 – 0.07 and gamma for 0.35 – 0.95. The models with a combination of 0.05 – 0.35 for learning rate and gamma had better accuracies of almost 89% whereas the rest combinations resulted into 86.7 – 87.8 %.

b. Observing graph for GCN with edge thresholds

We set two threshold values for two edge types E_a, E_k. If an edge between two nodes has a weight lower than the threshold value, then the edge will be removed. The resulting graph was used for training topic labels with a combination of 0.05 – 0.35 as the learning rate and gamma values.

The notable results were obtained from five threshold scenarios (Table 3) with some cases had no threshold (scenario 1, 2, 3 and 4). The TF-IDF threshold limits E_a (Eq. (3)) and the PMI threshold limits E_k (Eq. (4)). To balance the keyword number and still limit the thresholds while maintaining better accuracies, the recommended threshold combination is 0.2 for both edge types (scenario 5).

$$e_{a(a_i,k_j)} = \begin{cases} e_{a(a_i,k_j)} & \geq thres_{TF.IDF} \\ 0 & < thres_{TF.IDF} \end{cases} \quad (3)$$

$$e_{k(k_i,k_j)} = \begin{cases} e_{k(k_i,k_j)} & \geq thres_{PMI} \\ 0 & < thres_{PMI} \end{cases} \quad (4)$$

c. Observing GCN models with more topic labels

Our tests also checked the GCN performances for all 18 topics. Since the process of generating graph and representing feature matrix require some space, our framework suggested keyword selection based on DBSCAN clustering [16]. The original data after pre-process has ±109K short texts for 18 topics, while the empirical works of GCN employed ± 20-30K texts for four topics.

Some keywords that frequently occurred together in certain subjects tend to have similar contexts.

Table 4. Average performance values (in %) for article title classification on selected topics

Model		Acc.	Prec.	F1
Logistic Regression (LogRes)	TF-IDF	89.32	88.80	88.64
	GloVe	67.48	67.69	66.91
LSTM	TF-IDF	92.46	75.48	84.89
	GloVe	88.05	65.89	78.58
Bi-LSTM	TF-IDF	92.83	78.34	86.90
	GloVe	93.19	78.08	86.89
CNN	TF-IDF	93.09	77.43	80.42
	GloVe	88.14	70.11	80.42
The proposed framework (TF-IDF)		88.77	88.41	88.40

Visualization of those keywords into two dimensions by using t-SNE transformation will result in closely allocated points [17, 18]. The t-SNE strategy gives the benefit of simplifying high dimensional data into lower representation to support manual observation easily.

Our framework proposed t-SNE process on the data of ±109K and then employed DBSCAN clustering of those points with a cluster number of 18. The keywords with the same subject contexts would be seen as data points in a dense region. Therefore, to recognize those regions, the DBSCAN method is designated to accommodate the problem.

In our empirical analysis, two parameters of DBSCAN clustering were carefully observed: the distance threshold to include data points in clusters (*eps*) and the minimum number of data points to form a dense cluster (*min_samples*). The process obtained ± 20-30K texts for 18 topics. After processing those data into a graph and used for GCN training with 0.05 – 0.35 as the learning rate and gamma values, the accuracy was ± 68%. Our GCN model has considered the semantic context of keywords through a graph representation. Nevertheless, this third observation of accuracy results inferred that topic diversity (from 4 to 18 topics) had influenced the LDA label quality.

3.2 Evaluate classification performances

We evaluate the performances using accuracy, precision and harmonic mean F1-measure metrics. F-measure is a metric that combines precision and recall.

The empirical analysis on observing parameters of the GCN model concluded as follows: data from four topics (± 20K texts), edge thresholds of 0.2 for TF-IDF and PMI, in addition to GCN parameters of learning rate 0.05 and gamma value 0.35. We have synthesized a dataset from the original ± 209K of article texts into ± 20K. Then, the approach of Deep LDA in determining the topic correctness presented ± 81% accuracies (Table 2), whereas the approach of GCN or the proposed framework offered a higher value of ± 88% (Table 4). The results in Table 2 are also parts of the proposed framework for justifying label correctness and synthesizing sampling data.

As a comparison, the experiment results in Table 4 were implemented on other classification methods of a simple yet effective logistic regression as a starter [19], and the following deep learning-based classifiers: long short-term memory (LSTM) [20], Bi-LSTM, and convolutional neural network (CNN) [21].

Previous researches on short text classification [6-9] showed that deep learning are recommended. Thus, in this experiment we evaluated commonly utilized of deep learning architectures such as LSTM, Bi-LSTM, and CNN with a baseline comparison of non-deep learning approach of logistic regression. Table 4 displays the classification results using LDA topics as class labels. Since the word embedding usage based on frequency (TF-IDF) and context (GloVe) shows comparable results, our classification only considers TF-IDF as weighting scheme.

Preceding works with TF-IDF discussed keyword occurrences as word vector representation in data. Table 4 displays the results of methods using context to create word vector representation like GloVe [22], i.e., leveraging co-occurrences between keywords in the entire data. Here, the GloVe embedding ensures all word vector representation of ± 20K short texts are similarly distributed in a 300-dimensional space. The accuracies of other deep-learning methods with or without the embedding approach demonstrated better results than our proposed method. In Bi-LSTM case, the embedding usage ensures the stability in accuracies. However, their precision values are much lower. It means that some article titles of training data not included in generating the classifier model have a somewhat different context. The results in Table 4 are averaged from the topics of "law, education, computer engineering, and conservation".

As mentioned in Table 2, the "computer eng." topic has a relatively diverse keyword usage since the field of computer engineering might solve real problems related to law or education issues by implementing information and communications technology (ICT) techniques. Thus, the GCN advantage in relationship reserves between keywords could help subject context converge. However, this advantage is not supported by LogRes method, although it gave comparable results.

Our proposed framework has been tested to handle a dataset containing non-English short-texts

for classification without class labels with high possibility of overlapped context. Commonly used deep learning methods like LSTM, Bi-LSTM, and CNN that consider word sequences to represent subject context have displayed false positive with high accuracy low precision values. The graph usage that represents the word relations to overcome the overlapped context in deep learning (GCN) has displayed more balance state between accuracy and precision.

4. Conclusion and future works

This work proposes a framework for preparing article titles situated in a cold-start condition, which could be employed in a knowledge management system of researchers. Our empirical analysis started with diverse subjects has investigated the LDA-generated topic labels to help researchers in searching. Steps in the framework include preprocessing, word filtration, LDA process, and evaluation process using deep learning approaches with or without considering word relationships. The evaluation with commonly used deep learning methods has displayed relatively higher accuracies of a little more than 90%. However, the precisions were much lower since the relations between keywords could not be preserved. Our suggested process in the proposed framework successfully maintains those two performance indicators of around 89%. Furthermore, the empirical results showed that a more focused subject context is preferable for obtaining the appropriate topic labels even in a cold-start situation. The challenge is finding an optimal representation for generating and transforming text-keywords graphs into a feature matrix for more diverse subjects.

In the future, we will try to represent text-keywords into a visualization in virtual reality display that can be explored by human inspector (who wear VR HMD) and enable further explorations intuitively interactively.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Surya Sumpeno and Diana Purwitasari; *Methodology*, Diana Purwitasari; *Software*, Bastian Farandy and Dini Adni Navastara; *Validation*, Bastian Farandy and Surya Sumpeno; *Resources*, Diana Purwitasari; *Data curation*, Bastian Farandy; *Writing—original draft preparation*, Surya Sumpeno and Diana Purwitasari; *Writing—review and editing*, Surya Sumpeno and Mauridhi Hery Purnomo; *Visualization*, Dini Adni Navastara; *Supervision*, Surya Sumpeno and Mauridhi Hery Purnomo; *Project administration*, Diana Purwitasari.

Acknowledgments

This work was supported by Indonesian Ministry of Research and Technology under Grant No. 3/E1/KP.PTNBH/2021 with an institutional contract No. 932/PKS/ITS/2021.

References

[1] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. S. Petri, and S. Adam, "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology", *Communication Methods and Measures*, Vol. 12, No. 2–3, pp. 93–118, 2018.

[2] S. W. Kim and J. M. Gil, "Research Paper Classification Systems based on TF-IDF and LDA Schemes", *Human-Centric Computing and Information Sciences*, Vol. 1, No. 1, 2019.

[3] Y. Lee and J. Cho, "Web Document Classification using Topic Modeling based Document Ranking", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 11, No. 3, pp. 2386–2392, 2021.

[4] K. Kurata, Y. Miyata, E. Ishita, M. Yamamoto, F. Yang, and A. Iwase, "Analyzing Library and Information Science Full-Text Articles using a Topic Modeling Approach", In: *Proc. of the Association for Information Science and Technology*, Vol. 55, No. 1, pp. 847–848, 2018.

[5] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation", In: *Proc. of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Aoyama, Japan, pp. 165-174, 2017.

[6] D. Purwitasari, A. B. Ilmi, C. Fatichah, W. A. Fauzi, S. Sumpeno, and Mauridhi Hery Purnomo, "Conflict of Interest based Features for Expert Classification in Bibliographic Network", In: *Proc. of 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, Surabaya, Indonesia, pp. 54–59, 2018.

[7] T. T. Dien, N. T. Hai, and N. T. Nghe, "Deep Learning Approach for Automatic Topic Classification in An Online Submission System", *Advances in Science, Technology and Engineering Systems Journal*, Vol. 5, No. 4, pp. 700–709, 2020.

[8] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification", In: *Proc. the 33rd Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence,* Hawai, USA, pp. 7370-7377, 2019.

[9] K. Tayal, N. Rao, and K. Subbian, "Short Text Classification using Graph Convolutional Network", In: *Proc. the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019),* Vancouver, Canada, 2019.

[10] J. Xu, Y. Cai, X. Wu, X. Lei, Q. Huang, H. Leung, and Q. Li, "Incorporating Context-Relevant Concepts into Convolutional Neural Networks for Short Text Classification", *Neurocomputing*, Vol. 386, pp. 42–53, 2020.

[11] W. Cunha, V. Mangaravite, C. Gomes, S. Canuto, E. Resende, C. Nascimento, F. Viegas, C. França, W. S. Martins, J. M. Almeida, T. Rosa, L. Rocha, and M. A. Gonçalves, "On the Cost-Effectiveness of Neural and Non-Neural Approaches and Representations for Text Classification: A Comprehensive Comparative Study", *Information Processing and Management*, Vol. 58, No. 3, p. 102481, 2021.

[12] X. Qin, J. Liu, Y. Wang, Y. Liu, K. Deng, Y. Ma, K. Zou, L. Li, and X. Sun, "Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews", *Journal of Clinical Epidemiology*, Vol. 133, pp. 121–129, 2021.

[13] A. Ferligoj, L. Kronegger, F. Mali, T. A. Snijders, and P. Doreian, "Scientific Collaboration Dynamics in a National Scientific System," *Scientometrics*, Vol. 104, No. 3, pp. 985-1012, 2015.

[14] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", In: *Proc. the LREC 2010 Workshop on New Challenges for NLP Frameworks,* Valletta, Malta, 2010.

[15] M. R. Bhat, M. A. Kundroo, T. A. Tarray, and B. Agarwal, "Deep LDA : A New Way to Topic Model", *Journal of Information and Optimization Sciences*, Vol. 41, No. 3, pp. 823-834, 2020.

[16] S. Chen, X. Liu, J. Ma, S. Zhao and X. Hou, "Parameter Selection Algorithm of DBSCAN based on K-means Two Classification Algorithm", In: *Proc. the 7th International Symposium on Test Automation and Instrumentation (ISTAI),* 2018.

[17] L. J. P. V. D. Maaten and G. E. Hinton, "Visualizing High-Dimensional Data using t-SNE", *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605, 2008.

[18] D. Purwitasari, R. Alamsyah, D. A. Navastara, C. Fatichah, S. Sumpeno, and M. H. Purnomo, "Visualizing Academic Experts on a Subject Domain Map of Cartographic Alike", In: *Proc. the 4th International Conference on Computer, Communication and Computational Sciences (IC4S2019),* Bangkok, Thailand, 2019.

[19] I. Alsmadi and G. K. Hoon, "Term Weighting Scheme for Short-Text Classification: Twitter Corpuses", *Neural Computing and Applications*, Vol. 31, pp. 3819–3831, 2019.

[20] J. H. Wang, T. W. Liu, X. Luo, and L. Wang, "An LSTM Approach to Short Text Sentiment Classification with Word Embeddings", In: *Proc. the 2018 Conference on Computational Linguistics and Speech Processing*, Hsinchu, Taiwan, pp. 214-223, 2018.

[21] H. Wang, K. Tian, Z. Wu, and L. Wang, "A Short Text Classification Method based on Convolutional Neural Network and Semantic Extension", *International Journal of Computational Intelligence Systems*, Vol. 14, No. 1, pp. 367 - 375, 2021.

[22] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation", In: *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532–1543, 2014.