International Journal of Technology

http://ijtech.eng.ui.ac.id

*Research Article*

# Clustering Narrow-Domain Scientific Text Using Unsupervised and Similarity-Based Approaches

**Saiful Akbar [1,*], Anindya Prameswari Ekaputri [1], William Fu [1], Rahmah Khoirussyifa' Nurdini [1], Salman Ma'arif Achsien [1], Benhard Sitohang [1]**

[1]*School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jl. Ganesha 10, Bandung, 40132, Indonesia*

**Corresponding author: saiful@itb.ac.id, Tel.: +62222508135, Fax.: +62222508135*

**Abstract:** Clustering scientific papers published by authors is useful for discovering fellow authors with similar interests or research groups in the institution. In this study, we explore the use of scientific text clustering with an unsupervised approach to enhance the retrieval efficiency of similar works. Challenges in clustering scientific papers from a specific domain include an increase in the list of non-discriminating words (stop words) because more words are becoming common in most of the documents. For example, words such as *engineering* will no longer have discriminating power if most documents are from the engineering field. The use of similar terminologies to express different concepts, such as internet vs. internet of things, is also a challenge. To address this, we experimented with various text processing methods, including stemming, lemmatization, technical stop word removal, noun extraction, and n-gram phrase detection. The experiment was conducted on a corpus of faculty publications. Our methodology used text processing methods with latent Dirichlet allocation and non-negative matrix factorization topic models to cluster the documents and uncover latent topics within the corpus. The NMF model combined with lemmatization, technical stop word removal, noun extraction, and phrase detection was determined to be the optimal clustering pipeline. The pipeline yielded 11 clusters with the following evaluation scores: UMass of -2.493, CV of 0.681, NPMI of -0.136, and UCI of -4.491. It also improved the sample accuracy from 71.1% to 80.7% and generalized well to a different dataset. The resulting clusters from this pipeline fit our institution's research groups, such as electrical power engineering, signal processing, and computer vision. Additionally, we provide a curated list of technical stop words that contributed to the effectiveness of our clustering results.

**Keywords:** Latent dirichlet allocation; Narrow-domain Non-negative factorization matrix; Text clustering; Text processing; Topic modelling

## 1. Introduction

The increasing number of scientific publications provides more research ideas and acts as a reference for future innovations (Li et al., 2020; Larsen and Von Ins, 2010). However, it also raises the need for an efficient search process in handling a vast repository of texts. This is useful for upcoming authors to explore supporting works of literature for their future research and to discover fellow authors with similar interests (Sajid et al., 2021). To increase the quality of the search process, document clustering could be used to label a publication into the most suitable groups. The clustering results support retrieval processes to return more relevant documents (Zibani et al., 2022;

This work was supported by the School of Electrical Engineering and Informatics, Institut Teknologi Bandung, funded by P2MIGB Grant No. 968/IT1.C12/KU/2023

https://doi.org/10.14716/ijtech.v16i5.7110
Received May 2024; Revised June 2024; Accepted September 2024

Kadhim, 2019). Advancements in text analysis tasks, such as text classification (Aftab et al., 2023; Tey et al., 2023; Mohammed et al., 2021) and document grading (Lubis et al., 2021), are also reflected in text clustering (Mohammed et al., 2021), thus enabling the exploration of more possible cases.

Moreover, document clustering provides insight into documents within a collection, especially under unannotated conditions. Clustering research documents within an academic institution, even in one with established research groups, can help identify emerging research topics. Discovering topic groups enables the observation of research trends, which can help determine the research direction an institution is heading toward. Clustering can automate the management of the scientific publications archive as it can be used to assist in labeling documents automatically.

There have been notable approaches to clustering scientific documents in an institution (Pavithra and Savitha, 2024; Preetham et al., 2022; Bellaouar et al., 2021; Kim and Gil, 2019). The work presented in those papers aimed to discover groups of research interest existing within the faculty. Insights from the clustering process help faculty members learn about the research focus in the faculty. Furthermore, this insight gives references to the topics open for further development, thus igniting collaboration among faculty members.

Document clustering is a technique for grouping documents into clusters where documents in a cluster share common properties according to defined similarity measures (Shah and Mahajan, 2012). In contrast to document classification, where the number of clusters and the cluster for each document are known, document clustering does not include information about the number, characteristics, or members of the clusters. This makes document clustering a type of unsupervised learning.

Topic modeling is an approach to cluster documents and discover useful topics from each cluster (Muchene and Safari, 2021; Vayansky and Kumar, 2020). Recently, word embeddings have been used to vectorize document contents, and then the documents are grouped based on their similarity in the vector space (Mehta et al., 2021). Other approaches to discover insights from text documents include co-word analysis (Leung et al., 2017; Surjandari et al., 2015). Among these methods, the most commonly used approaches for topic modeling tasks are latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) (Smail et al., 2023; Yu and Xiang, 2023; Mifrah and Benlahmar, 2020).

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative model that identifies topics from a collection of text documents. This model assumes that a document is a mixture of topics, and a topic could be viewed as a distribution of words. The LDA learning process aims to discover the word distribution within a topic. This is achieved by calculating the probability of a word given the topic and updating the distribution of probability for each topic. LDA has been used in various ways, such as topic modeling in linguistic science, political science, biomedical fields, geographical locations, and social networks (Jelodar et al., 2019). In context of scientific document clustering has been used to classify documents in environmental education (Chang et al., 2021) and profile publications related to Industry 4.0 (Janmaijaya et al., 2021).

Non-negative matrix factorization (NMF) (Lee and Seung, 1999) is another topic modeling approach that utilizes a non-negative matrix structure of dimension m x n as the document-term matrix. In this method, a matrix V is approximated as a product of two matrices W and H with the dimensions of m x k and k x n. In topic modeling, k represents the number of topics discovered from the corpus. Thus, the weights on each vector column on W represent the rank of words within the topic, and the weight on each column of H represent the proportions of topics within that document. NMF is typically used for dimensional reduction and clustering (Hassani et al., 2021; Wang and Zhang, 2013; Tsuge et al., 2001), and several studies have specifically utilized NMF to cluster documents (Laxmi Lydia et al., 2020; Shahnaz et al., 2006). LDA and NMF perform differently on short documents, with the latter considered better at discovering distinct topics (Egger and Yu, 2022).

In this study, we applied topic modeling techniques to discover research interest groups in the dataset of publications from the School of Electrical Engineering and Informatics (STEI) Institut

Teknologi Bandung. The nature of STEI publications, which focused on the field of computer science and electrical engineering, raised the difficulty of clustering the documents due to an increase in noise and very similar texts.

There is an increase in noise because more words appeared in most, if not all, of the documents. For a word to have a discriminating value, it must be different. It must appear uniquely on few documents that share some similarity. A word that appears in most documents cannot become an identifying feature to determine the document cluster; thus, it becomes noise to the dataset. These words are called stop words. The challenge of increasing the number of stop words in domain-specific texts was addressed in a previous study (Sarica and Luo, 2021). The research focused on a broader range of engineering. In this paper, we focused on more specific fields with the hope of producing a more effective collection of technical stop words, especially in the computer science and electrical engineering domains.

Having texts from a very specific domain of engineering also caused the texts were very similar because similar terminologies were used to explain different concepts. For example, in the context of machine learning, the word "model" refers to an algorithm trained on data (e.g., neural network model), whereas in the context of software design, the word "model" refers to an abstract representation of a system (e.g., UML model). This ambiguity was addressed by grouping words that describe a concept into phrases, with the hope that phrases can better capture the nuance of the discussion.

Before proceeding to document clustering, we experimented with several text processing methods, including the process of removing domain-specific stop words and grouping words to phrases, to determine the most optimal pipeline for scientific text clustering. Each experiment was evaluated with coherence scores as quality metrics, as done in existing works (Hadiat, 2022; Mifrah and Benlahmar, 2020). Finally, we observed and evaluated a sample of our results to gain more understanding of the clusters formed.

The contributions of this study are as follows:

1. Curated a list of stop words related to computer science and electrical engineering. Our corpus, which consists of titles and abstracts of computer science and electrical engineering publications, was analyzed to identify words with no discriminative value (stop words) in the computer science and electrical engineering domain.
2. Experimentation to combine various text preprocessing steps. Various combinations of text preprocessing steps were compared for discovering topic clusters within scientific publications. The preprocessing methods we experimented include stemming and lemmatization, technical stop-word removal, noun extraction, and n-gram phrase detection.
3. Finally, we proposed an optimum text clustering pipeline that is most suited to our text clustering task based on our observation of our experiment results.

The paper consists of several sections as follows: The first section describes the background for these experiments, including a quick summary of previous works that support our topic. The second section provides a description of the case study dataset along with our proposed experiment design. The experimental results are displayed and analyzed in the following sections. Finally, the last section contains all citations in the manuscript and presents all additional discoveries from the experiments.

## 2. Methods

### *2.1. Data*

We used publications from the School of Electrical Engineering and Informatics (STEI), Institut Teknologi Bandung as our dataset for this case study. This dataset contains the title, authors, publication year, and abstract of scientific publications authored by the lecturers of the institution. This dataset is compiled from the STEI owned dataset and various digital publications platforms, including Google Scholar and IEEE Xplore. The data were collected from November to December 2022 with 10246 titles.

After preliminary data cleaning, 10246 unique publication titles were available. However, determining and removing duplicates present some challenges. One example is among publications that are determined to be duplicated, some possess identical titles but are published in different venues and thus count as two separate publications. Manual or semi-manual checking using a script may be required for further data cleaning. Citing this complexity, analysis is done without removing presumed duplicates for the sake of this manuscript. We obtained the publication year for most (8777 out of 10246) of the collected publications. As shown in Figure 1, most of the publications collected were published after the start of 2006, which coincides with the formation of STEI. Publications dated before 2006 were mostly books authored by senior lecturers.

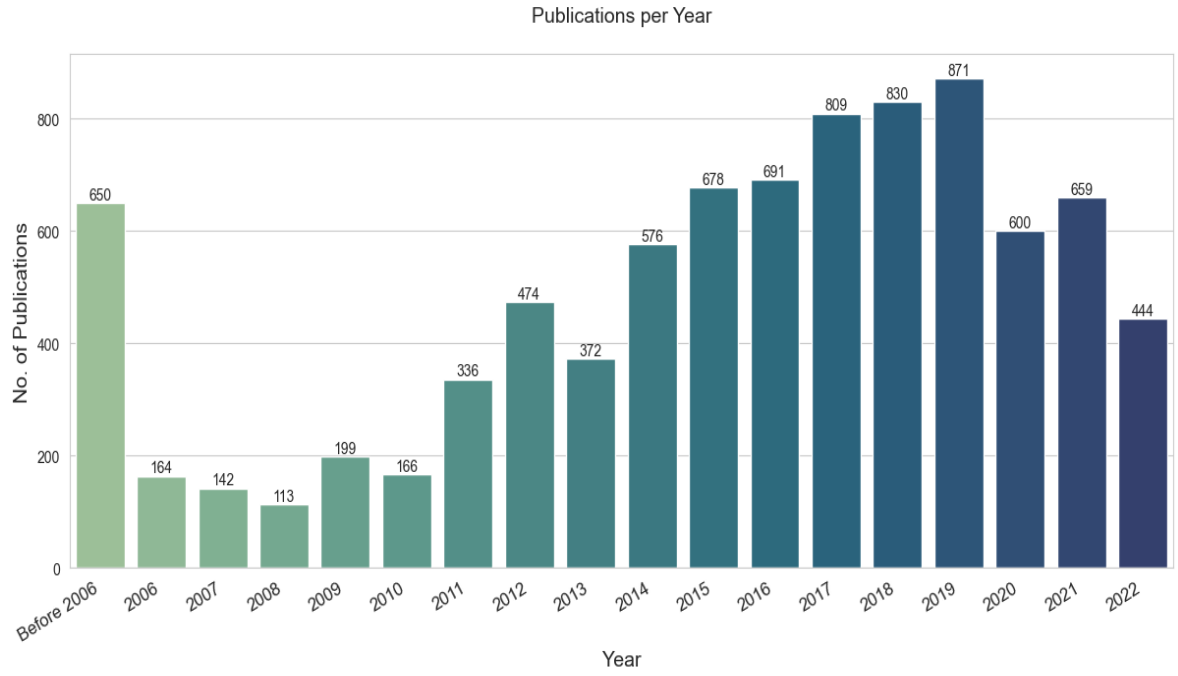


**Figure 1** Distribution of publications produced by STEI faculty members by year

Most of the collected publications are in English (9330 papers), with a smaller portion being in Indonesian (905 papers) and other languages, such as Japanese (3 papers), German (3 papers) and others (5 papers). Each lecturer in the dataset belonged to a specific research group within the faculty. Among the 10246 collected, we notice the participating research groups for 9767 publications: the remaining 479 are blank, most likely due to the author's name being mislabeled because of different spellings. We also discovered that some publications were part of more than one research group. For example, a publication on the Internet of Things (IoT) is part of the collaboration between lecturers from the EL and IF research groups. By counting the number of entries containing the code for each research group, the number of publications in which each research group contributed was obtained, as shown in Figure 2.

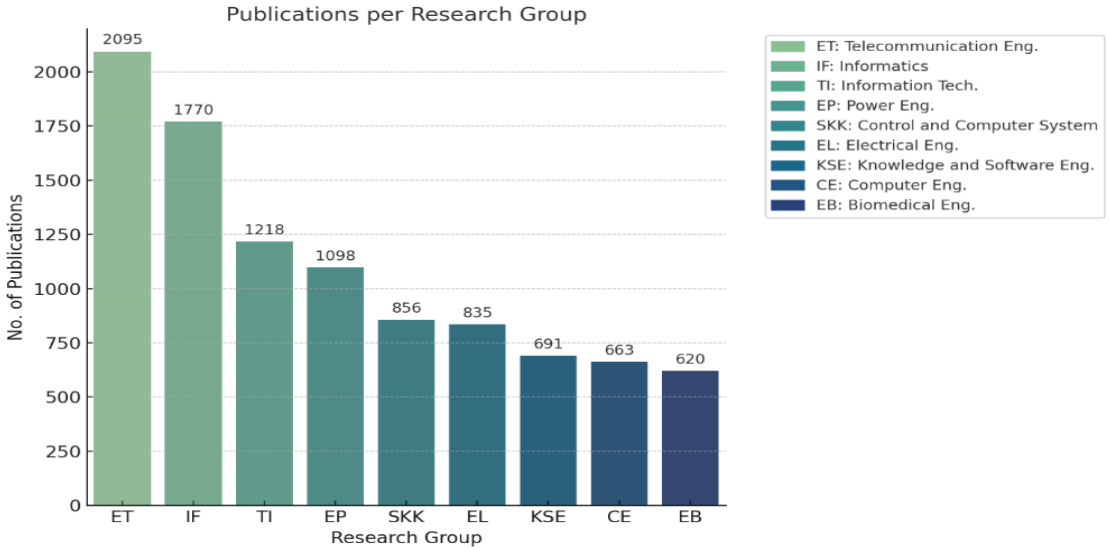


**Figure 2** Number of publications produced by STEI faculty members across different research groups

*2.2. Experiments*

In this section, we explain our experiment scenario. The general steps for our document clustering flow are shown in Figure 3.

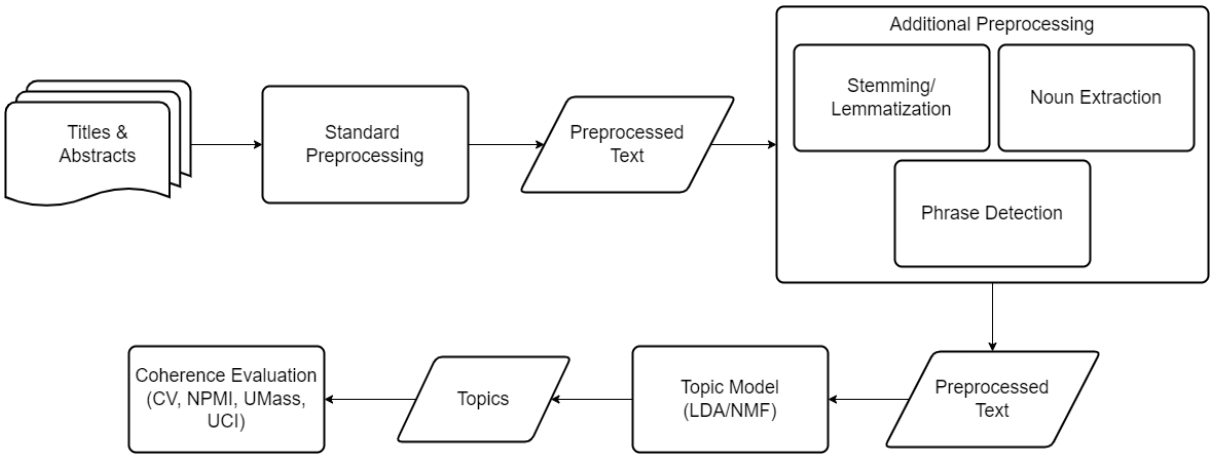


**Figure 3** The initial process of the experiment

In terms of scientific papers, we use only titles and abstracts (Kim and Gil, 2019) or full-text documents (Terko et al., 2019) to perform clustering. Research (Syed and Spruit, 2017) stated that a collection of scientific articles from a specific domain would perform better with full-text data, whereas articles from a broader range would perform better with only the abstracts. Since our data contains articles from two main knowledge areas (informatics and electrical engineering), we decided to use only the titles and abstracts.

Subsequently, we performed basic preprocessing to the texts: lowercase, removing punctuations and numbers, removing basic stop words (with stop words as listed in the NLTK library), translating non-English documents, and removing duplicate entries. There are also non-paper publications in the corpus, for example, committee documents. As we were unsure of the importance of those documents, we decided to make two versions of our corpus: one with the non-paper documents included (non-filtered) and one with the non-paper documents excluded (filtered). We tested the models on both versions.

An experiment is then conducted to obtain the most optimal preprocessing pipeline. This experiment aims to discover which preprocessing methods could improve performance and which methods hurt performance instead. We tested three preprocessing methods, as summarized in Table 1. Each of these methods will be explained in detail in the following sections.

**Table 1** Preprocessing steps for the evaluation

| Name | Process Objective |
|---|---|
| Stemming, lemmatization, and removal of technical stop words | Removing noises and ensuring that words with the same meaning have the same representation |
| Noun extraction | Capturing only important entities and simplifying data |
| Phrase detection | Grouping concepts "electrical engineering" vs. "electrical" "engineering" |

For each method, we performed a grid search (with doing method vs without doing method) and varied the types of documents (filtered vs unfiltered) and the model type (LDA vs NMF), which were then tested with coherence scores. The best methods are then included in the pipeline to build the final model, and we will further analyze its results.

#### *2.2.1. Standard Preprocessing*

For each method, we performed a grid search (with doing method vs without doing method) and varied the types of documents (filtered vs unfiltered) and the model type (LDA vs NMF), which were then tested with coherence scores. The best methods are then included in the pipeline to build the final model, and we will further analyze its results.

#### *2.2.2. Pipelining Experiment*

For each experiment, we compared the performance of the LDA and NMF of the two topic modeling methods against the filtered and unfiltered corpus. We used LDA and NMF models because they are less computationally expensive than DL methods and BERTopic (Grootendorst, 2022). As our corpus dataset is relatively small and consists of short texts, the use of data-hungry BERT (Devlin et al., 2018) algorithms may be less effective for this case study. Using LDA and NMF models also allows us to further analyze the formed cluster and gain more information about our dataset, as opposed to DL methods that use a black box approach (Zini and Awad, 2023).

Table 2 shows the parameters set for this experiment. We set our initial number of topics as 9 to match the number of research groups in the faculty. Each experiment was evaluated by its coherence scores, namely, CV, UMass, NPMI, and UCI. We also examined the top final models' topic frequency and the top 5 words for each topic.

**Table 2** Set the parameters for the experiments

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| n_topics: | 9 | Iterations / max_iter: | 10 000 |
| random_state: | 42 | passes: | 5 |

**Experiment 1: Preprocessing.** This experiment aims to evaluate which preprocessing pipeline has the best effect on our cluster results. Three kinds of preprocessing methods were tested: lemmatizing, stemming, and technical stop word removal.

Lemmatizing is turning the word into its root form, for example, *writing* and *written* becomes *write*. Stemming is cutting the word into a short form, for example, *writing* and *written* becomes *writ*. Both stemming and lemmatization aim to ensure that words with the same meaning do not have different representations. As in the previous example, writing and written should not be represented differently. Stemming is usually faster than lemmatization because lemmatization requires a dictionary lookup, but stemming might cause confusion and ambiguity because different

words are cut into the same representation. For example, stemming *universal* and *university* would produce the same representation (*univers*) despite their different meanings. Lemmatization would correctly differentiate universal and university but would require a longer time due to the dictionary lookup process. We conducted an experiment to determine whether it is better to use lemmatization or stemming for our corpus.

In contrast to basic stop words, technical stop words are words that frequently occur in a domain. Stop words are words that are removed before any text processing is executed because they are insignificant and do not add any meaning (Rajaraman and Ullman, 2011). In general, stop words are filler words such as *am, is, are, he*, and *she*. In this case study, our corpus mainly contains scientific papers, so there are new stop words relating to academia. For example, the words *methods, data*, and *paper* are generally not considered stop words. However, because our corpus is more specific toward academic publications, those words will appear in almost every document as they relate to experiment methods. Thus, those words became stop words because they do not add significant information about the text they appeared on. These new stop words will be referred to as "tech stop words."

Detection of technical stop words has been researched (Sarica and Luo, 2021). In the research, they analyzed texts in the patent database to identify stop words in the engineering domain based on statistical measures. The final stop word list was constructed by creating sorted lists based on the statistical measures, which were then evaluated by humans. The final list consists of 26 new stop words, combined with 62 words from a previous study, to produce a list of 87 technical stop words.

After constructing the list, a case study on multi-class classification with LSTM was conducted (Sarica and Luo, 2021). The results showed that removing technical stop words increased the precision, recall, and accuracy scores compared with models trained with raw texts and with texts that had only basic stop words removed.

In this experiment, a grid search was performed on lemmatization vs. stemming the words and with vs. without technical stop words. Lemmatizing and stemming were performed using the NLTK library. For the technical stop words, we used the strategy used in (Sarica and Luo, 2021) to identify the stop words in our corpus. We did not use the final list because our corpus included texts from the electrical engineering department, which may have a different set of stop words than the one used in the manuscript (Sarica and Luo, 2021). There are 4 combinations from the grid, multiplied by the number of models (2) and corpus (2). In this experiment, we tested a total of 16 models.

**Experiment 2: Noun extraction.** This experiment aims to determine whether using full text or just noun phrases to cluster our documents is better. This is based on Kim and Gil (Kim and Gil, 2019), who removed stop words and extracted only nouns to reduce the number of processed texts and improve processing efficiency. We applied the same method to our experiment because the most important words and concepts in our domain were nouns, e.g., internet of things, machine learning, robotics, and signal processing. For the implementation, we used SpaCy and Liamca's noun phrase extraction algorithm (https://github.com/liamca/noun-phrase-extraction).

We also varied the phrasing method, checking whether it is better to group the phrased nouns (*New York* becomes "*new_york*") or separate them (*New York* becomes "*new*" and "*york*"). In this experiment, we would like to discover the significance of bigram or trigram phrase detection to our topic models, since these detections might provide detailed topic words. For example, the term *neural network* is widely used in deep learning-based research; however, when a document is considered as a singular token, either *neural* or *network* term independently would give a different context.

**Experiment 3: Phrase detection.** This experiment aimed to determine whether it is better to process phrases as phrases or separate words. For example, considering biomedical engineering as one concept (biomedical_engineering) instead of separate words (biomedical and engineering)

might be better. These words sequences are often called n-grams, where n is the number of words in a sequence. The phrase *biomedical_engineering* is considered a bigram (2 g).

N-gram detection for clustering texts was researched by Mohemad et al. (Mohemad et al., 2021). In the study, they clustered crime event-related texts to group them into five classes of modus operandi (MO). They used phrase detection as a preprocessing step and experimented with three variations: 2, 3, and 4 g. The results showed that detecting 2-gram and 3-gram did not improve the results, but 4-gram had the best performance, even exceeding the baseline.

In this study, we experimented with three types of n-grams: 1-gram, 2-gram, and 3-gram. Although Mohemad et al. (Mohemad et al., 2021) stated that 4-gram has the best performance, we decided to only experiment with 1, 2, and 3-grams because 2 and 3-grams were more commonly used, while 4-gram usage is quite rare. Also, Mohemad et al. (Mohemad et al., 2021) used data from a completely different domain, so it is safer to experiment on commonly used n-grams than on rare ones.

**Experiment 4: Final model**. After obtaining the best variations from each experiment, we conducted one more experiment to build a stronger model based on previous results. We will note which parameters from the experiment give better results and then use those parameters in our fourth experiment.

After building the model, we conducted an elbow method to test whether there is a more fitting number of topics other than 9 because a paper in one research group could be further divided into more specific domains. We also looked into the keywords for each topic in the optimal model for further analysis.

## 3. Results and Discussion

### *3.1. Preliminary Experiment*

The best results for Experiments 1, 2, and 3 are summarized in Table 3, and the full results are presented in the supplementary file of this paper.

**Table 3** Result summary for experiment 1-3

| No | Parameter | CV | NPMI | UMass | UCI |
|---|---|---|---|---|---|
| 1 | NMF + filtered + lemma + tech stop words removed | 0.554 | 0.062 | -2.618 | 0.113 |
| 2.a | NMF + unfiltered + phrased | 0.689 | -0.134 | -2.493 | -4.589 |
| 2.b | NMF + unfiltered + unphrased | 0.679 | 0.113 | -2.421 | 0.705 |
| 3 | LDA + unfiltered + 1-gram | 0.494 | 0.047 | -1.851 | 0.316 |

Experiments 1 and 2 showed that NMF is the best method for clustering our corpus, whereas experiment 3 performed best with LDA, but with a far lower score. Reference (Egger and Yu, 2022) stated that when dealing with shorter texts, NMF is better than LDA. As our texts consist of only the title and abstracts, not the entire publication text, NMF is suitable because of the corpus's short length. Experiments 2 and 3 showed that using unfiltered texts (including non-scientific documents) is better, whereas experiment 1 showed that the filtered corpus is better. We further investigated this by building two models, one with filtered and the other with unfiltered text, and comparing the results. We then analyzed the results for each experiment step as follows.

**Additional Preprocessing.** From Experiment 1, lemmatization performed better than stemming. This is because stemming is prone to ambiguity. As previously explained, stemming reduced *universal* and *university* to the same representation *universe* despite the two words having different meanings, whereas lemmatization represented the two words correctly because of the dictionary look-up process. Our corpus has many similar words with different meanings, where stemming those words causes more ambiguity, making it difficult for the model to cluster the documents correctly. To remove technical stop words, we applied the algorithm from Salica and Luo (Sarica and Luo, 2021) to our corpus. This approach produced 88 technical stop words, the

complete list of which is displayed in Appendix B. The study (Sarica and Luo, 2021) produced a similar number of new stop words (87); however, their list consisted of more general words (mentioned, accordingly, furthermore, instead), whereas our list consisted of words related to academia (process, performance, result, improve, technique, data). Despite having more specific words, stop word removal successfully improved our model scores, indicating that the identified words were truly words that add no additional meaning, and removing them reduces the ambiguity caused by those words. From this experiment, we decided that lemmatization and tech stop words removal would be included in our final model pipeline.

**Phrase Detection.** Based on Experiment 3, the model with 1 g performs better than the model with 2 and 3 g. After running the phrase detection function, we investigated the most frequent terms in our corpus and found that the top-20 terms were mostly still in the form of 1-gram, with a frequency of at least 1500. The top-1 term, *system*, appeared approximately 8000 times. The most frequent 2-gram, *real_time*, is far behind by appearing only 571 times, while the most frequent 3-gram, *inspect_non_controlled*, is even more far behind with only 160 appearances. This proves that despite detecting the phrases, they still appeared less frequently than the commonly used 1-gram terms. As both LDA and NMF used word count (bag-of-words) in their algorithm, terms that appeared less would have less weight to determine the topic to which a document belongs. This means that even though 2 and 3-grams were detected, they do not have much impact on improving model performance because they do not appear as frequently as 1-grams, hence having less weight compared to the 1-grams. Detecting 2 and 3 gram increased the number of new stop words. Among the common 2 and 3-grams, we found phrases that do not add new information about the document topic, for example the phrases *paper_presents, case_study, design_implementation, proposed_method*, and *result_show*. Moreover, these stop phrases appeared more often than the meaningful ones, so phrase detection adds more noise to the corpus. This could be improved by first removing the technical stop words and then running the phrase detection algorithm. However, we concluded that this would not be as effective because of the low n-gram appearance problem, which was previously explained. Thus, we decided to exclude phrase detection from the final model pipeline.

**Noun Extraction.** The model built with a corpus that consisted of nouns only produced the highest CV scores, indicating that noun extraction was the most helpful among all preprocessing we experimented on. This proves that the most important words and concepts in our corpus are indeed nouns; thus, eliminating non-noun words removes noise and allows the model to focus only on important features. Noun extraction worked great, so we included noun extraction in our final model pipeline. However, there is little difference in CV score between phrased and unphrased nouns. Similar to the case in phrase detection, this could be because common phrases do not appear as often as common words (2 and 3-gram appeared less than 1-grams), so grouping phrases might have no significant difference on how the documents were clustered. Despite having similar CV scores, the difference in UCI scores was drastic: the phrased corpus UCI score was significantly lower than the score for the unphrased corpus. We decided to further investigate this issue by experimenting on both phrased and unphrased corpus when building our final model and later compared their topic results.

### *3.2. Final Model*

Based on experiment 1-3, NMF, 1-gram, lemmatization, and technical stop words removal produced the best results. However, phrased vs. unphrased words were tied, and experiment 1 showed that it is best to use filtered texts, while experiment 2-3 showed that using unfiltered texts is better. Due to the score ties and inconsistency, an additional experiment was conducted for our final model. We ran a grid search with filtered text vs. unfiltered text and phrased vs. unphrased texts; four final models were tested in total. We also conducted elbow methods to determine the optimal number of clusters for all four models. The results are summarized in Table 4.

**Table 4** Result summary for Experiment 4

| No | phrased | filtered | n_cluster | UMass | CV | NPMI | UCI |
|---|---|---|---|---|---|---|---|
| 1. a | yes | no | 9 | -2.493 | 0.689 | -0.134 | -4.589 |
| 1. b | yes | no | 11 | -2.493 | 0.681 | -0.136 | -4.491 |
| 2 | yes | yes | 18 | -2.605 | 0.652 | -0.120 | -4.337 |
| 3 | no | no | 9 | -2.421 | 0.679 | 0.113 | 0.705 |
| 4 | no | yes | 13 | -2.724 | 0.653 | 0.103 | 0.510 |

From the table above, we noticed that the filtered corpus needs a higher number of topics. This probably occurred because the filtered corpus excluded general documents such as conference committees and preface texts, leaving only technical documents and publications. Those documents are more specific to a domain; thus, more topics are needed to properly cluster them. We will further analyze the topics by comparison as explained below.

We concluded that the best clustering model is constructed by NMF with lemmatized, stop word removed, phrased, and unfiltered corpus. It is best constructed using nine topics. For more details about the results, we further analyzed Model 1a as it produced the best numerical results. Close seconds, 1.b and 3, were also analyzed for comparisons. First, the topic words for model 1a are shown in Table 5. The nine research groups in our department can be divided into two larger groups: electrical engineering (EL) and informatics (IF). The topics produced by Model 1a showed more topics from EL (topics 1, 6, 7, 8, and 9) than those from IF (topics 2, 4, and 5). This is natural as there are more research groups in EL than in IF. There is also a unique topic that combines renewable energy from EL and green IT from IF (Topic 3), probably because both topics discuss the environment. Several topics have large distributions (topics 1, 7, and 8) exceeding 1300. This probably happens because those topics are more general than the other more specific topics, such as NLP (Topic 4), e-learning (Topic 5), or antenna (Topic 9). Those big topics probably consisted of more diverse documents, and publications that did not fit into the more specific topic were likely classified into those big topics.

We chose n_topics = 9 because there are nine research groups; however, the clustering results do not match our existing research groups. Several topics from a research group appeared more than once. For example, topics 1 and 7 are both from the electronics research group, while topics 6 and 9 are from telecommunication engineering. There is no cluster representing topics from biomedical engineering or electrical power engineering. This could happen due to the imbalance in the number of publications in each research group. For example, biomedical engineering is relatively new, so there might not be as many publications from the research group. Another reason for this is that some research groups cover more topics than others. For example, electronics could also cover some basic electrical power engineering and control systems and computers.

Next, we compared those topics with those produced by Model 1b. Table 6 shows the list of topics. The model produced two more topics than Model 1a, but the numerical scores for both models are similar. Compared to topics from Model 1a, we found several new topics, namely, electrical power (1b.2), signal processing/fiber optics (1b.5), and computer vision (1b.7). The topic that appeared in 1. but not in 1b is renewable energy/green IT (1a.3).

The new topics might be formed due to the higher number of topics, so the model could classify the documents into more specific clusters. Hence, no big clusters with distributions over 1300 exist. However, there is one very small cluster that consists of only 467 instances (1b.1, IoT), which is interesting because its counterpart from 1a (1a.8, IoT) is the largest cluster. This confirms that cluster 1a.8 consists of various documents that are slightly unrelated to IoT, which can then be broken down into several new topics.

**Table 5** Topic from Model 1a

| ID | Topic Words | Interpreted Topic | Number of Documents Required |
|---|---|---|---|
| 1a.1 | output, input, current, motor, speed, controller, vehicle, experimental, low, component, motor, motor, speed | robotics | 1421 |
| 1a.2 | management, framework, business, organization, case, government, activity, architecture, concept, and important | IT enterprise and governance | 819 |
| 1a.3 | energy, case, load, renewable, source, cost, electricity, generation, potential, plant | renewable energy/green IT | 1008 |
| 1a.4 | indonesian, classification, best, word, machine, text, language, extraction, sentence, vector, machine, machine, language, extraction, vector | NLP | 811 |
| 1a.5 | learning, student, architecture, education, processing, activity, teacher, machine, medium, game | e-learning | 874 |
| 1a.6 | rate, error, channel, parameter, bit, scheme, term, noise, low, wireless | wireless communication | 1148 |
| 1a.7 | voltage, characteristic, parameter, effect, experimental, current, material, discharge, partial, property [Remark 1] | electrical/material experiments | 1357 |
| 1a.8 | device, internet, function, thing, sensor, human, mobile, smart, protocol, main | IoT/ smart device | 1449 |
| 1a.9 | substrate, dielectric, antenna, epoxy, characterization, fr4, dimension, thickness, structure, microstrip, characterization, | antenna | 731 |

Similarly, Topic 1a.3 might disappear because the instances grouped in 1a.3 found a better-suited cluster in Model 1b. Thus, the cluster related to green IT grows bigger (1b.10, IT enterprise/governance with 1030 documents) than its counterpart (1a.2, IT enterprise/governance with 819 documents). On the other hand, instances related to renewable energy were grouped into cluster 1b.2 (electrical power), which is a new topic. This clustering also made sense because finding renewable energy is a common research topic for power generation.

Thus, despite having similar metric scores, Model 1b appeared to produce better clustering results by human judgment. However, Model 1b still does not include some minor research groups, such as biomedical engineering or computer engineering, which is reasonable because the number of publications from those research groups is far smaller than that from the other research groups.

These results were also compared with our second runner-up model, which is Model 3. This model also produced 9 topics but was conducted without phrasing the nouns. The scores for Model 3 are similar to those of models 1a and 1.b, except for the exceptionally low UCI. Table 7 presents the topics from Model 3.

Despite having the same number of topics as Model 1a, Model 3 produced almost entirely different clusters. The only recurring topics with those produced in 1.a are electrical systems (3.2), NLP (3.4), and IT enterprise/governance (3.8). Interestingly, a wide gap exists in the distribution of these topics. For example, Topic 3.8 has almost 2200 instances, whereas Topic 3.2 has only 492 instances. Meanwhile, the newly appearing topics seem to be a more specific version of those mentioned in 1a and 1b, while some others are a mash-up of several topics in 1a and 1b.

**Table 6** Topics from Model 1b

| ID | Topic Words | Interpreted Topic | Number of Documents Required |
|---|---|---|---|
| 1b.1 | sensor, environment, monitoring, important, thing, internet, iot, mobile, dynamic, smart | IoT/smart device | 467 |
| 1b.2 | current, energy, voltage, load, output, electric, experimental, source, inverter, renewable energy | electrical power | 1133 |
| 1b.3 | learning, machine, classification, word, best, language, text, extraction, neural | NLP | 1080 |
| 1b.4 | characteristic, voltage, discharge, effect, material, insulation, pattern, important | electrical/material | 949 |
| 1b.5 | low, range, light, circuit, rate, standard, noise, receiver, modulation, source, standard, noise, standard | signal processing/fiber optics | 839 |
| 1b.6 | function, controller, linear, speed, solution, position, cost, motor, error, platform | robotics | 631 |
| 1b.7 | processing, object, device, part, computer, architecture, digital, field, human, camera | computer vision | 696 |
| 1b.8 | learning, student, activity, education, internet, concept, medium, digital, experience, interaction, student | e-learning | 968 |
| 1b.9 | substrate, dielectric, antenna, epoxy, characterization, fr4, dimension, structure, thickness, microstrip, characterization, | antenna | 589 |
| 1b.10 | case, management, framework, business, organization, government, solution, existing, tool, framework, existing | IT enterprise and governance | 1030 |
| 1b.11 | parameter, error, rate, channel, wireless, scheme, evaluation, access, transmission, term, transmission | wireless communication | 1236 |

For example, Topic 3.1 shared several keywords with Topic 3.9. However, Topic 3.1 focused more on antennas and their design elements (proven with the keywords *radar, substrate, epoxy*, and *patch*), while Topic 3.9 focused more on high-frequency systems (proven with the keywords *filter, waveguide*, and *microstrip*). It might be better to combine those two topics as they have subtle differences and discuss antenna / signal. Moreover, the two topics also have a smaller instance count compared to other clusters, proving that the two were indeed very specific.

In contrast to topics 3.1 and 3.9, which became very specific, we also found several topics that mashed up several domains in a cluster. Topic 3.5 seems to be clustering smart devices and robotics together, whereas they were separated in models 1a and 1b. This cluster still made sense because both smart devices and robotics have several aspects that collide, e.g., both need sensors and interact with the environment. The more erroneous cluster happened on Topic 3.7, where it seems to combine e-learning and ML. Although both concepts include the word "learning," they were two entirely different topics. Documents about e-learning were mostly about software engineering and constructing new applications, while machine learning is about finding patterns in data and predicting patterns in unseen data. These erroneous clusters might be the reason why the UCI score for Model 3 increased significantly compared with the UCI score for models 1a and 1b.

Based on the above analysis, we concluded that Model 1b, with phrased nouns, unfiltered documents, and 11 topics, was the best model for clustering lecturers' publications in STEI.

**Table 7** Topics from Model 3

| ID | Topic Words | Interpreted Topic | Number of Documents Required |
|---|---|---|---|
| 3.1 | dielectric, substrate, structure, filter, epoxy, waveguide, fr4, response, microstrip, bandwidth, ethyl ether, epoxy, | microwave / high-frequency systems | 664 |
| 3.2 | voltage, discharge, partial, oil, insulation, transformer, characteristic, electric, electrical, current | electrical systems | 492 |
| 3.3 | energy, current, load, renewable, electric, source, solar, plant, motor, hybrid, renewable, electric | renewable energy | 700 |
| 3.4 | classification, language, text, word, extraction, sentence, recognition, speech, based | NLP | 1079 |
| 3.5 | smart, device, sensor, mobile, monitoring, home, environment, protocol, platform, robot | smart devices/robotics | 1206 |
| 3.6 | channel, low, low, low, error, estimation, voltage, scheme, output, input, controller | communication systems | 1912 |
| 3.7 | learning, student, game, machine, education, environment, mobile, activity, language, deep, language | e-learning / gamification / mobile | 902 |
| 3.8 | management, framework, architecture, business, government, digital, case, organization, enterprise, and engineering | IT enterprise and governance | 2182 |
| 3.9 | antenna, array, patch, substrate, microstrip, radiation, radar, gain, bandwidth, epoxy | antenna/radar | 481 |

### *3.3. Result Validation*

We sampled 200 random publications from our STEI dataset and manually labeled them to the topic group from our model that we believe the publication belongs to. We then ran the clustering pipeline to the two versions of the sample: the raw data and the data after performing noun detection, phrasing, and technical stop words removal. For fair comparison, we performed lowercase, punctuation, and general stop word removal, as well as lemmatization for both versions. We evaluated the results, which are shown in Table 8.

Performing noun detection, phrasing, and technical stop words increases the sample's accuracy from 71.1% to 80.7%. In comparison, a previous study (Sarica and Luo, 2021) applied stop-word removal for a clustering task and reached an accuracy of 95.9% and 97.0% for datasets with general and technical stop words removed, respectively. A summary of this comparison is shown in Table 8. While the previous research achieved higher accuracy, our approach presented more improvements from the baseline, proving that noun detection, phrasing, and technical stop words removal can effectively improve clustering performance for scientific texts.

**Table 8** Metric comparison between this study and previous research

| Evaluated Texts | Precision | Recall | Accuracy |
|---|---|---|---|
| **This research** | | | |
| Removed general stop words | 0.710 | 0.715 | 0.711 |
| Noun detection, phrasing, and removal of general and technical stop words | 0.815 | 0.807 | 0.807 |
| (Sarica and Luo, 2021) | | | |
| Removing the general stop word | 0.961 | 0.959 | 0.959 |
| Removal of general and technical stop words | 0.971 | 0.970 | 0.970 |

We also performed our pipeline on a sampled arXiv dataset. For this experiment, we randomly sampled 100 papers from the computer science and 100 papers from the category of electrical engineering and systems science to mimic the nature of our dataset. We also limited the time frame to include only papers published from 2020 to 2024.

To compare the performance of our dataset, we used Hellinger distance, which measures the similarity between documents from the same topic and other topics. This metric calculates the distance between two probability distributions, making it suitable for topic vectors of LDA or NMF topic models (Muchene and Safari, 2021). The Hellinger similarity between topics was computed by calculating the distance between every document pair of different topics and averaged their scores. The Hellinger similarity is the complement of the Hellinger distance, whose value falls between 0 and 1. Two topics will be considered similar when the Hellinger similarity is close to zero, and vice versa. The results are shown in Figure 4.

From this metric, we observed that documents within each group have few overlapping topics. The relatively lower similarity score between document groups shows this. We noticed that documents in the same group also had a relatively stronger similarity score, although its value is relatively far from a strong similarity, which is closer to one. This observation is similar for both our STEI and the sampled arXiv text documents.

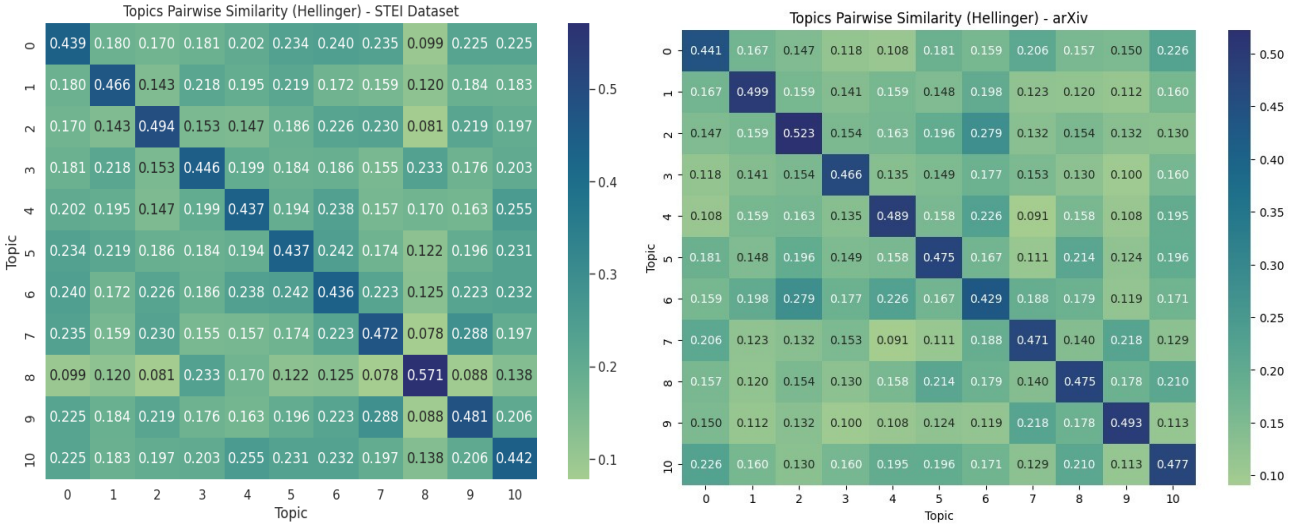


**Figure 4** Topics Pairwise Similarity using Hellinger distance for the STEI (left) and arXiv (right) datasets

## 4. Conclusions

In this study, we explored combinations of several text processing methods while addressing the challenge of clustering texts from specific domains: increased technical stop words and the use of similar terms for different concepts. Our findings indicate that the Non-Negative Matrix Factorization (NMF) model combined with lemmatization, technical stop word removal, noun extraction, and phrase detection performed best among all the combinations we tested. This model effectively grouped our institution's scientific documents into 11 clusters and improved the sample clustering accuracy from 71.1% to 80.7%. We have also applied it to another dataset and confirmed that our method is generalizable, as the results were similar to those obtained when our dataset was applied. The results have practical implications for improving the effectiveness of literature retrieval, discovering emerging research trends within an institute, and automating literature labelling. However, our study focused on a corpus from a single institution, specifically in electrical engineering and informatics, which may limit its applicability to other domains or interdisciplinary studies. Future research could explore the application of this methodology to a broader range of scientific fields other than engineering.

## Acknowledgements

This work was supported by the P2MIGB Grant No. 968/IT1.C12/KU/2023 from the ITB School of Electrical Engineering and Informatics.

## Author Contributions

The authors confirm their individual contributions as follows: Saiful Akbar contributed to the conceptualization, manuscript drafting, review, editing; Anindya Prameswari Ekaputri and William Fu were involved in manuscript drafting, literature review, and experimentation; Rahmah Khoirussyifa' Nurdini and, Salman Ma'arif Achsien were responsible for data collection and experimentation. Benhard Sitohang was involved in conceptualization and review. All authors have read and agreed to the published version of the manuscript.

## Conflict of Interest

The authors have no conflicts of interest to declare.

## References

Aftab, F, Bazai, SU, Marjan, S, Baloch, L, Aslam, S, Amphawan, A & Neo, TK 2023, 'A comprehensive survey on sentiment analysis techniques', *International Journal of Technology*, vol. 14, no. 6, pp. 1288-1298, https://doi.org/10.14716/ijtech.v14i6.6632

Bellaouar, S, Bellaouar, MM & Ghada, IE 2021, 'Topic modeling: Comparison of LSA and LDA on scientific publications', *In:* Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering, pp. 59–64, https://doi.org/10.1145/3456146.3456156

Blei, DM, Ng, AY & Jordan, MI 2003, 'Latent Dirichlet allocation', *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022

Chang, I-C, Yu, T-K, Chang, Y-J & Yu, T-Y 2021, 'Applying text mining, clustering analysis, and latent Dirichlet allocation techniques for topic classification of environmental education journals', *Sustainability*, vol. 13, no. 19, article 10856, https://doi.org/10.3390/su131910856

Devlin, J, Chang, M-W, Lee, K & Toutanova, K 2018, 'BERT: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint, http://arxiv.org/abs/1810.04805

Egger, R & Yu, J 2022, 'A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts', *Frontiers in Sociology*, vol. 7, https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M 2022, 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', arXiv preprint, http://arxiv.org/abs/2203.05794

Hadiat, AR 2022, 'Topic modeling evaluations: The relationship between coherency and accuracy', Thesis, University of Groningen, viewed 01 August 2023, (https://fse.studenttheses.ub.rug.nl/28618/1/s2863685_alfiuddin_hadiat_CCS_thesis.pdf)

Hassani, A, Iranmanesh, A & Mansouri, N 2021, 'Text mining using nonnegative matrix factorization and latent semantic analysis', *Neural Computing and Applications*, vol. 33, no. 20, pp. 13745–13766, https://doi.org/10.1007/s00521-021-06014-6

Janmaijaya, M, Shukla, AK, Muhuri, PK & Abraham, A 2021, 'Industry 4.0: Latent Dirichlet allocation and clustering based theme identification of bibliography', *Engineering Applications of Artificial Intelligence*, vol. 103, article 104280, https://doi.org/10.1016/j.engappai.2021.104280

Jelodar, H, Wang, Y, Yuan, C, Feng, X, Jiang, X, Li, Y & Zhao, L 2019, 'Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey', *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169-15211, https://doi.org/10.1007/s11042-018-6894-4

Kadhim, AI 2019, 'Survey on supervised machine learning techniques for automatic text classification', *Artificial Intelligence Review*, vol. 52, no. 1, pp. 273-292, https://doi.org/10.1007/s10462-018-09677-1

Kim, S-W & Gil, J-M 2019, 'Research paper classification systems based on TF-IDF and LDA schemes', *Human-Centric Computing and Information Sciences*, vol. 9, no. 1, article 30, https://doi.org/10.1186/s13673-019-0192-7

Larsen, PO & von Ins, M 2010, 'The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index', *Scientometrics*, vol. 84, no. 3, pp. 575-603, https://doi.org/10.1007/s11192-010-0202-z

Laxmi Lydia, E, Krishna Kumar, P, Shankar, K, Lakshmanaprabu, SK, Vidhyavathi, RM & Maseleno, A 2020, 'Charismatic document clustering through novel K-means non-negative matrix factorization (KNMF) algorithm using key phrase extraction', *International Journal of Parallel Programming*, vol. 48, no. 3, pp. 496-514, https://doi.org/10.1007/s10766-018-0591-9

Lee, DD & Seung, HS 1999, 'Learning the parts of objects by non-negative matrix factorization', *Nature*, vol. 401, pp. 788–791, https://doi.org/10.1038/44565

Leung, XY, Sun, J & Bai, B 2017, 'Bibliometrics of social media research: A co-citation and co-word analysis', *International Journal of Hospitality Management*, vol. 66, pp. 35-45, https://doi.org/10.1016/j.ijhm.2017.06.012

Li, Y, Wang, K, Xiao, Y & Froyd, JE 2020, 'Research and trends in STEM education: A systematic review of journal publications', *International Journal of STEM Education*, vol. 7, no. 1, article 11, https://doi.org/10.1186/s40594-020-00207-6

Lubis, FF, Mutaqin, Putri, A, Waskita, D, Sulistyaningtyas, T, Arman, AA & Rosmansyah, Y 2021, 'Automated short-answer grading using semantic similarity based on word embedding', *International Journal of Technology*, vol. 12, no. 3, pp. 571-581, https://doi.org/10.14716/ijtech.v12i3.4651

Mehta, V, Bawa, S & Singh, J 2021, 'WEClustering: Word embeddings based text clustering technique for large datasets', *Complex & Intelligent Systems*, vol. 7, no. 6, pp. 3211-3224, https://doi.org/10.1007/s40747-021-00512-9

Mifrah, S & Benlahmar, EH 2020, 'Topic modeling coherence: A comparative study between LDA and NMF models using COVID-19 corpus', *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5756-5761, https://doi.org/10.30534/ijatcse/2020/231942020

Mohammed, SM, Jacksi, K & Zeebaree, RM 2021, 'A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms', *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, article 552, https://doi.org/10.11591/ijeecs.v22.i1.pp552-562

Mohemad, R, Muhait, NNM, Noor, NMM & Othman, ZA 2021, 'The impact of N-gram on the Malay text document clustering', *Malaysian Journal of Information and Communication Technology*, vol. 6, no. 2, pp. 22-29, https://doi.org/10.53840/myjict6-2-83

Muchene, L & Safari, W 2021, 'Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya', *PLOS ONE*, vol. 16, no. 1, article e0243208, https://doi.org/10.1371/journal.pone.0243208

Pavithra & Savitha 2024, 'Topic modeling for evolving textual data using LDA, HDP, NMF, BERTopic, and DTM with a focus on research papers', *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2, pp. 53-63, https://doi.org/10.37802/joti.v5i2.618

Preetham, MCS, Reddy, BR, Tharun Reddy, DS & Gupta, D 2022, 'Comparative analysis of research papers categorization using LDA and NMF approaches', *In:* Proceedings of the 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), pp. 1-7, https://doi.org/10.1109/NKCon56289.2022.10127059

Rajaraman, A & Ullman, J 2011, 'Data mining', in *Mining of Massive Datasets*, Cambridge University Press, pp. 1–17, https://doi.org/10.1017/CBO9781139058452.002

Sajid, NA, Ahmad, M, Afzal, MT & Atta-ur-Rahman 2021, 'Exploiting papers' reference's section for multi-label computer science research papers' classification', *Journal of Information & Knowledge Management*, vol. 20, no. 1, article 2150004, https://doi.org/10.1142/S0219649221500040

Sarica, S & Luo, J 2021, 'Stopwords in technical language processing', *PLOS ONE*, vol. 16, no. 8, article e0315195, https://doi.org/10.1371/journal.pone.0254937

Shah, N & Mahajan, S 2012, 'Document clustering: A detailed review', *International Journal of Applied Information Systems*, vol. 4, no. 5, pp. 30-38, https://d1wqtxts1xzle7.cloudfront.net/81705889/ijais12-450691-libre.pdf

Shahnaz, F, Berry, MW, Pauca, VP & Plemmons, RJ 2006, 'Document clustering using nonnegative matrix factorization', *Information Processing and Management*, vol. 42, no. 2, pp. 373-386, https://doi.org/10.1016/j.ipm.2004.11.005

Smail, B, Aliane, H & Abdeldjalil, O 2023, 'Using an explicit query and a topic model for scientific article recommendation', *Education and Information Technologies*, vol. 28, no. 12, pp. 15657-15670, https://doi.org/10.1007/s10639-023-11817-2

Surjandari, I, Dhini, A, Wibisana, N & Lumbantobing, EWI 2015, 'University research theme mapping: A co-word analysis of scientific publications', *International Journal of Technology*, vol. 6, no. 3, pp. 410-421, https://doi.org/10.14716/ijtech.v6i3.1462

Syed, S & Spruit, M 2017, 'Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation', *In:* Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA 2017), pp. 165-174, https://doi.org/10.1109/DSAA.2017.61

Terko, A, Zunic, E & Donko, D 2019, 'NeurIPS conference papers classification based on topic modeling', *In:* Proceedings of the 2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT), pp. 1-5, https://doi.org/10.1109/ICAT47117.2019.8938961

Tey, WL, Goh, HN, Lim, AHL & Phang, CK 2023, 'Pre- and post-depressive detection using deep learning and textual-based features', *International Journal of Technology*, vol. 14, no. 6, pp. 1334-1343, https://doi.org/10.14716/ijtech.v14i6.6648

Tsuge, S, Shishibori, M, Kuroiwa, S & Kita, K 2001, 'Dimensionality reduction using non-negative matrix factorization for information retrieval', *In:* Proceedings of the 2001 IEEE International Conference on Systems, Man and Cybernetics, pp. 960-965, https://doi.org/10.1109/ICSMC.2001.973042

Vayansky, I & Kumar, SAP 2020, 'A review of topic modeling methods', *Information Systems*, vol. 94, article 101582, https://doi.org/10.1016/j.is.2020.101582

Wang, Y-X & Zhang, Y-J 2013, 'Nonnegative matrix factorization: A comprehensive review', *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, https://doi.org/10.1109/TKDE.2012.51

Yu, D & Xiang, B 2023, 'Discovering topics and trends in the field of artificial intelligence: Using LDA topic modeling', *Expert Systems with Applications*, vol. 225, article 120114, https://doi.org/10.1016/j.eswa.2023.120114

Zibani, P, Rajkoomar, M & Naicker, N 2022, 'A systematic review of faculty research repositories at higher education institutions', *Digital Library Perspectives*, vol. 38, no. 2, pp. 237-248, https://doi.org/10.1108/DLP-04-2021-0035

Zini, JE & Awad, M 2023, 'On the explainability of natural language processing deep models', *ACM Computing Surveys*, vol. 55, no. 5, pp. 1-31, https://doi.org/10.1145/3529755