Global Transitions Proceedings 3 (2022) 298–304

Contents lists available at ScienceDirect

Global Transitions Proceedings

journal homepage: http://www.keaipublishing.com/en/journals/global-transitions-proceedings/

Prediction of research trends using LDA based topic modeling

Rahul Kumar Gupta*, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, Thoudam Doren Singh

Department of Computer Science and Engineering, NIT Silchar, Assam, India

ARTICLE INFO

Keywords:
Applied intelligence
LDA
Probabilistic approach
TF-IDF
Topic modeling

ABSTRACT

Change is the only constant. In many sectors, a change is being witnessed that is getting increasingly rapid. This carries a plethora of new innovation possibilities with it. This necessitates well-founded data about trends, future developments and their consequences. This study seeks to catch the new directions, paradigms as predictors with an association of each topic which will be discovered through topic modeling techniques like LDA with BoW. For this, empirical analysis on 3269 research articles from the Journal of Applied Intelligence was done, which were gathered during a 30-year span. The inferred topics were then structured into a way suitable for performing predictive analysis. This is significant in the sense that it will help to predict what technology will be encountered in the future, as well as how far human's ability to innovate and discover things may lead this world to. The final model using TF-IDF scores has outperformed the baseline model by a margin of 41%.

1. Introduction

Research is the key to the advancement of mankind. It is a process of discovering a new domain of knowledge. Hence, it is of utmost importance that the researchers and all other stakeholders put their resources and energy into the specific area that shows a scope for development. This demands the need for quantitative and predictive analysis of various research trends [1]. The needs for which are data, their empirical analysis and a good predictive analysis of trends in domains [2,3]. The abbreviations used are listed in Appendix A.

1.1. Problem definition

With the expansion of the internet throughout the world it's tough to get relevant and required information with the expanding volume of data in recent years, most of which is unstructured. Nevertheless, this has made different technologies take center stage. Topic modeling [4] is one such approach in the field of text classification. It is a process which identifies the topic of a text corpus. The most popular technique is the **Latent Dirichlet Allocation** (LDA) [5–7]. Based on the words it contains, its goal is to determine which topic a document belongs to. Thus, this work involves the analysis of research trends in 3269 articles published under **"Applied Intelligence"** from 1991 to 2021, using the application of LDA [8–10]. The results can be used to further predict the future growth of different domains of Applied Intelligence using some statistical models [11–13]. A statistical model which captures the linear relationship between multiple entities over a time period will be an ideal one [14,15]. As a future part of the research the results of LDA can be fed into the selected forecasting model to do predictive analytics on future trends [16].

Centralizing the thoughts towards the area of research, organizations and geeks are required to put in their merits towards the notable domains of research. Diminishing the breadth of domains will lead to better upcoming revolutions in technologies. Some common hassles in predicting research trends include their innumerable methodologies [17] and falling short of the precisions as there is no touchstone with which the results can be compared with. Therefore, a two-way methodology where the analysis is being done not only using normalized abstract of the collected articles but also using the keywords as well has been put forward. The equation which governs the working of LDA is shown in Eq.(1).

$$P(W,\ Z,\ \theta,\ \varphi;\ \alpha,\ \beta) = \prod_{j=1}^{M} P(\theta_j;\ \alpha) \prod_{i=1}^{K} P(\varphi_i;\ \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\ \varphi_{j,t}) \tag{1}$$

Where, P(W, Z, θ, φ; α, β) is the probability that a specific document is produced from a hypothetical LDA machine; the settings of which will aid in knowing the latent topics in the document. α and β are the Dirichlet while θ and φ are Multinomial distributions respectively; Z is the list of topics and W is the corpus of Words. The value of P relies on four independent probabilities as mentioned in the right side of the Eq.(1) where K is one of the three model hyper parameters, i.e. number of topics, M and N denotes number of documents and word count in a given document respectively.

* Corresponding author.
E-mail address: rahulkr_ug@cse.nits.ac.in (R.K. Gupta).

https://doi.org/10.1016/j.gltp.2022.03.015

Available online 2 April 2022

1.2. Contribution

After completing the topic modeling stage, generation of word clouds will be done using the output of the LDA model which will be used for topic labeling. At this stage, the intersection of the keywords collected from the article database and the outcomes of the model will be used to further purify the results. These topics may be later used for forecasting and analysis [18,19].

The inculcation of TF-IDF scores of each word in the corpus before feeding it to the LDA model will be beneficial in obtaining accurate latent topics in the documents [20,21].

2. Related Works

The research trend prediction is carried out based on several techniques but more prominently based on machine learning techniques. Some of the most relevant literature surveys in conformation with the current topic are discussed below.

Using LDA, Gatti et al. [22] proposed a work on historical analysis of the Field of OR/MS using topic models in 2015.There are many possible directions for further investigation of the dataset used herein and the model created. It has an accuracy score of 0.87 but this study is specific to the OR/MS field. Using topic modeling, abstract parsing, rhetorical function labeling, a paper was published by Prabhakaran et al. [23] in 2016 on predicting the rise and fall of scientific topics from trends in their rhetorical framing which is a novel framework for assigning rhetorical functions to associations between scientific topics and papers. The accuracy score of the same is found to be 0.86. This framework assigns rhetorical functions to associate between scientific topics and papers. Nevertheless, examining only 20 years of scientific progress limits the framework from analyzing drastic scientific changes. Choi et. al. [24] in the year 2017 presented a work on analyzing research trends in personal information privacy using LDA topic modeling which provides results that are more comprehensive than those qualitative reviews in-depth which are above mentioned in the introduction. Krenn et al [25] predicted research trends with semantic and neural networks with an application in quantum physics using neural networks, NLP and network theory in the year 2019 with an accuracy score of 0.85. Singh et al. [26] in 2020 used PMI Score, TF-IDF, Text Graph Convolutional Neural Network, LDA, and proposed a work on hybrid classification with an accuracy score of 0.87. Kouassi et al. [27] in 2020 presented a work on analysis of deep neural networks for predicting trends in time series data using deep neural networks, TreNet, ensemble methods. Sivanandham et al. [28] used LDA topic modeling, NLP, time series analysis, vector auto regression for analyzing research trends using topic modeling and trend Prediction in 2021. Although the methods are clearly explained, the dataset spans only for articles around 9 years which may not produce actual results. Lee et al. [29] in June, 2021 applied deep neural networks, deep learning, t-SNE algorithm, BERT, time series analysis for future prediction growth potential of technologies with an accuracy score of 0.87. In this literature, ten promising technologies were identified. The criteria applied to the selection of the technology clusters are not absolute; thus, the selected technology clusters could be changed if a different criterion is chosen.

3. Method

The proceedings for the design of the system are appropriately described in this section. The complete flow diagram where the system could be depicted is shown in Fig. 1(a).

3.1. Data collection

The articles' useful data viz., date of publication, title, author details and abstract of the article have been extracted.

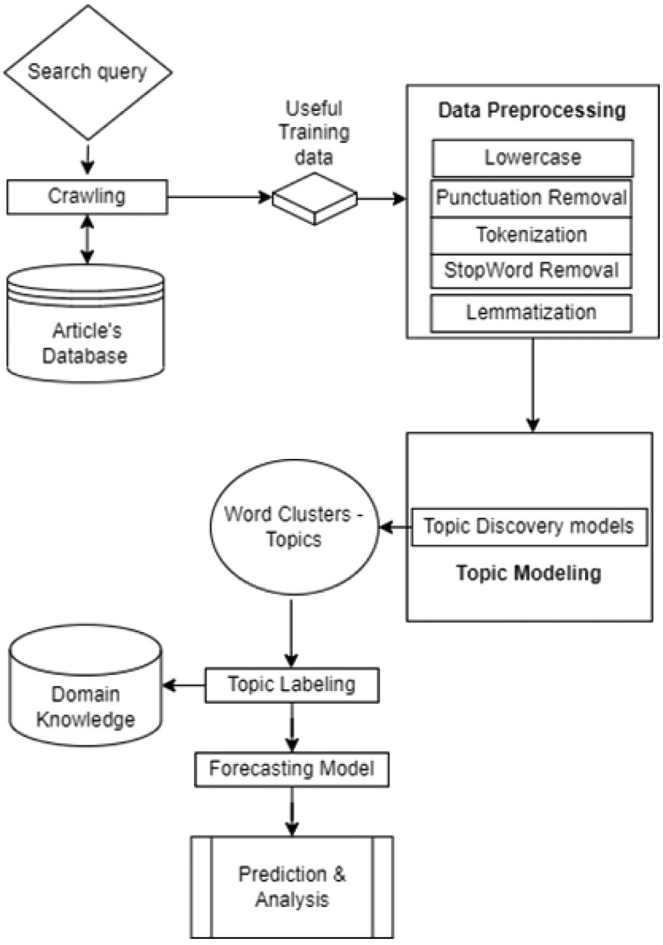

Fig. 1. (a). Flow Diagram of system **Fig. 1(b).** Blueprint of Proposed Model.

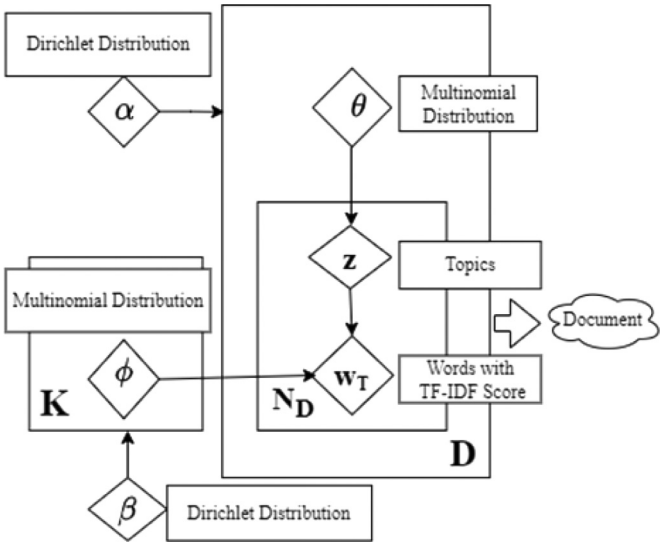

Fig. 1. Continued

For this purpose the technique of web crawling was employed to crawl over the entire database of this journal between the chosen timeline and the details were stored in a CSV (comma separated values) file.

```
#Baseline LDA Model
lda_model_base = gensim.models.LdaMulticore(bow_corpus, num_topics=10,
                                            id2word=dictionary,
                                            passes=5, workers=4)
```

Fig. 2. Running Baseline Model.

3.2. Pre-processing

An unstructured data usually contains a lot of irrelevant information which should be removed before actually passing on the data for training and analysis. This stage involves many steps, but here tokenization, stop word removal and lemmatization techniques for preprocessing of abstracts of the articles are followed.

Lowercase: Collected data is transformed into an uniform case, i.e. lowercase

Punctuation Removal: There is a need to remove the punctuation because some of the words embedding models don't support them.

Tokenization: Tokenization is a way of alienating bits of text into smaller units which are referred to as tokens.

Stop Word Removal: These are basically the words that don't add much value to the semantics of the content. After tokenization, this must have to be done which has the following benefits:

1.Size of the dataset eventually decreases after the removal of stop words which also leads to reduced training time.

2.Elimination of these inessentials will help to boost the performance as the space is narrowed down.

Lemmatization: This step reduces the inflected words. It aids in fetching the valid and necessary words.

3.3. Topic modeling

Basically, Topic modeling [30] is an unsupervised machine learning strategy which has the ability of checking a set of documents, identifying words and uncover the patterns within them and consequently cluster word bunches and comparative expressions that best characterize a set of corpus. In short, topic modeling algorithms generate collections of phrases and words that they think are related, allowing the reader to understand what those relationships mean, while classifying topics.

In this work, the LDA model is employed as it works well because it expresses competing sparsity between the topic distribution of documents(α) and the word distribution of topics(β). Inherently, LDA is a probabilistic approach to topic modelling. In other words, it is a Bayesian hierarchical probability generation model for collecting discrete data which works based on the assumption of compatibility between words and topics in the document. It models the corpus as a discrete distribution of the entire subject, and later subjects are displayed as the discrete distribution of the entire term in the document. It is an advance over the other models that also uses competing sparsity because the balance is resolved at the level of documents. This technique was developed by David Blei, Andrew Ng, and Michael Jordan and exposed in Blei et al., 2003 [2].

In this paper, the implementation of LDA is done though Gensim, which is a python library used for topic modeling, document indexing and similarity retrieval with large corpus.

3.3.1. Baseline LDA model using Bag of Words (BoW)

LDA is an excellent tool for covertly distributing topics in a large corpus which have the ability to identify subtopics for technical domains made up of multiple patents, where each patent is represented in an array of subject distributions. Here, a document is considered to be a mixture of subjects, where subjects are probability distributions for a set of terms where each document is considered a probability distribution over the set of topics. The LDA model is created in Gensim by simply specifying the corpus the dictionary mapping, and number of topics to be used in the model. Fig. 2 shows the execution of the Baseline model using BoW.

BoW is basically a corpus containing the word id and its frequency in each document. BoW corpus is created from a list of documents and text files. Formation of BoW corpus is shown in Fig. 3 and a sample output is depicted in Fig. 4.

3.3.2. Hyper parameter tuning

It is also known as Automatic model tuning which finds the best version of a model by running multiple tasks, which test a set of hyper parameters on the dataset. The influence of the LDA model's alpha and beta hyper parameters on the features of the baseline model is discussed in the following section.

Alpha is a parameter that controls the pre-distribution of the weights of topics in each document, while beta is the parameter for the pre-distribution of the weights of words in each topic. High alpha means every document is likely to contain a mixture of most of the topics and not just any single topic specifically, whereas low alpha means a document is more likely to be represented by just a few of the topics [34].

High beta means each topic is likely to contain a mixture of most of the words, not just any word specifically whereas low beta means the topic may contain a mixture of just a few words.

3.3.3. Proposed model: LDA utilizing TF-IDF Corpus

Fig. 1(b) shows the blueprint of the proposed model, where symbols carry usual meanings and $\boldsymbol{W_T}$ is the TF-IDF Corpus to be discussed in detail in this section. After finding out the appropriate model parameters, there is a need to imply some changes in the baseline model which may gradually increase the previously obtained results. In addition to the three hyper parameters, this research work tries to inculcate the rectitude of TF-IDF score which may be calculated as shown in Fig. 5.

TF-IDF is also a bag of words paradigm but is different from the normal corpus because it weighs the tokens i.e. words that appear often many times in the corpus. During initialization, it requires a training dataset with integer values like a BoW model which has already been obtained. In other words, in order to get a TF-IDF score first there is a need to train the corpus and then apply that corpus within the aforementioned model. To calculate the score, Gensim's *Tfidf-Model* has been used.

These scores can be very useful for topic modeling as it can be employed to visualize topics or rather better, to choose the vocabulary. This is supported by the fact that it is computationally costly to always use the entire dictionary. Having a look into the equation, Eq. (2) defining TF-IDF score will eventually make things more sensible about how it actually purifies the vocabulary which refers to the dictionary in this work.

$$\mathrm{tfidf(i,\ d, D)\ =\ tf(t,\ d) * idf(t,\ D)} \tag{2}$$

Where *tfidf* is the score to be calculated, tf refers to term frequency and is given by Eq. (3):

$$\mathrm{tf(t, d)\ =\ \frac{f(t, d)}{\sum_{t' \in d} f(t', d)}} \tag{3}$$

$f(t, d)$ is the frequency of a term t in a single document d. *idf* is the inverse document frequency and is given by Eq.(4):

$$idf(t, D) = \log \frac{|D|}{|\ \{\ d \in D\ :\ t \in d\ \}\ |} \tag{4}$$

```
bow_corpus = [dictionary.doc2bow(doc) for doc in processed_docs]
print("Number of BOW : ",len(bow_corpus))
# the corpus that contains the word id and its frequency in each document
bow_corpus[0]
```

Fig. 3. Creating BoW corpus.

```
Word 0 ("adopt") appears 1 time.
Word 8 ("better") appears 1 time.
Word 15 ("compare") appears 1 time.
Word 20 ("different") appears 2 time.
Word 23 ("experiment") appears 1 time.
Word 28 ("include") appears 1 time.
Word 32 ("information") appears 3 time.
Word 35 ("learn") appears 1 time.
```

Fig. 4. Visualization of BoW.

```
from gensim import corpora, models
tfidf = models.TfidfModel(bow_corpus)
corpus_tfidf = tfidf[bow_corpus]
from pprint import pprint
for doc in corpus_tfidf:
    pprint(doc)
    break
```

Fig. 5. Calculation of TF-IDF Scores.

Numerator denotes the total number of documents while denominator gives the number of documents in D such that the term t appears at least once in d. Apparently, if there exists a term t' such that it appears in all the documents then the *idf* score will be 0 and so will be the *tfidf* value making t' an elimination from TF-IDF corpus. Hence, in this way without even interfering with the basics of LDA it attempts to purify the vocabulary which directly has an impact on the results.

The basic steps involved in the proposed model are discussed here in the form of an algorithm.

1. Mapping normalized words and their integer ids on the preprocessed data
2. Converting document into a list of indices and replacing all the undefined words
3. Filtering out extreme tokens in the dictionary by their frequencies
4. Creating a BoW, this generates a vocabulary of all the unique words arising in all the documents.

Calculating TF-IDF score

Hyper parameter tuning, to set the right combination of hyper parameters which allows the model to maximize its performance.

Finding Coherence score, this evaluates a unique topic by measuring the semantic similarity between words that score high in the topic.

Word Cloud generation

4. Results

The average similarity between top words in a topic with the highest weights i.e. relative distance between the top words is measured by Coherence Score. This score can be used in topic modeling to measure how people interpret the topics. This is critical when the generated subject is used to view user-provided document collections or to understand trends and advancements in a certain field of study. The Gensim library has a *CoherenceModel* class which is used to find the coherence score of the proposed model.

For the baseline model, where no hyper parameter tuning or external efforts were made into the standard LDA modeling and only the BoW

```
# Compute Coherence Score for baseline model
coherence_model_lda = CoherenceModel(model=lda_model_base,
                                     texts=processed_docs,
                                     dictionary=dictionary,
                                     coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)

Coherence Score:  0.34169551572258944
```

Fig. 6. Training of Baseline Model.

```
# Compute Coherence Score for model using TF-IDF
coherence_model_lda = CoherenceModel(model=lda_model_tfidf,
                                     texts=processed_docs,
                                     dictionary=dictionary,
                                     coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)

Coherence Score:  0.3984663815376815
```

Fig. 7. Training of TF-IDF Model.

Table 1
1.(a). Tuning on 75% corpus Table 1(b). Tuning on 100% corpus.

Topics	Alpha	Beta	Coherence	Topics	Alpha	Beta	Coherence
2	0.01	0.01	0.436950	2	0.01	0.01	0.280892
2	0.01	0.31	0.251688	2	0.01	0.31	0.375246
2	0.01	0.61	0.417119	2	0.01	0.61	0.274719
2	0.01	0.9099999	0.278194	2	0.01	0.9099999	0.278838
2	0.01	symmetric	0.372537	2	0.01	symmetric	0.384062

Max Coherence Score: 0.47535
Max Coherence Score: 0.43877

was used for the purpose; the coherence score came out to be 0.342 as shown in Fig. 6. Making use of the already prepared TF-IDF corpus as an input to the LDA model with everything else remaining the same and the score improved to 0.398(Fig. 7).

Hyper parameter tuning was carried out on 75% and 100% corpus to find out the best model parameters. A sample view of output is shown in Table 1(a) and Table 1(b), along with the values of parameters corresponding to maximal coherence score. In Fig. 8, a scatter plot for Number of Topic vs Coherence score is shown which helped in selection of the value of k to be 7. K couldn't be chosen as 4 because it is not well-scattered. So, now the focus spans around k=7. For which a sample view is as shown in Table 2.

Finally, $\alpha = 0.01$ and $\beta = 0.90999999$ were chosen, which performed well and generated a score of nearly 0.408, however the score was increased to 0.483 by utilizing TF-IDF corpus instead of BoW and increasing the number of passes to 10, which is something that most of the works in this domain have failed to achieve (Fig. 9).

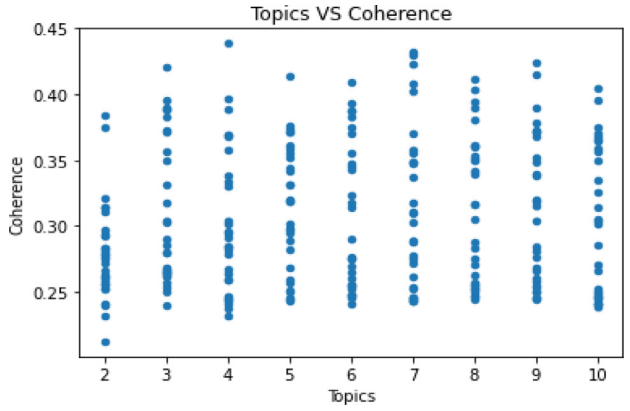

Fig. 8. Scatter Plot.

Table 2
Choosing Number of Topics.

Validation_Set	Topics	Alpha	Beta	Coherence
100% Corpus	7	0.01	0.01	0.347127
100% Corpus	7	0.01	0.31	0.370683
100% Corpus	7	0.01	0.61	0.317794
100% Corpus	7	0.01	0.9099999	0.408247
100% Corpus	7	0.01	symmetric	0.354909

5. Discussion

Seven topics were discovered after running the final model with the appropriate model parameters, each represented by a set of words with weights given to them. An attempt was made to find which single topic was more dominant in each of the documents, as shown in Table 3.

In this work, word clouds have been employed to present the seven topics as identified by the model. This type of raw display will help readers to get apt interpretations. However, the most taken way of representing latent topics is through a set of representative words (fixed numbers) and their respective weights. The word clouds for a particular compile time have been shown in Fig. 10.

Table 4 provides a comparison by quoting all of the current state-of-the-art, which will clearly differentiate the proposed model and make the fact glassy about where this model stands out among all of the listed literatures.

6. Conclusion

The impact of computationally modeling the evolution of science, by tracking how scientific topics rise and fall over time, on research funding and public policy is significant. The mechanisms underlying topic growth and decline, on the other hand, are poorly understood. The primary challenge addressed in this work is whether it is feasible to forecast the future development potential of technologies using data from relevant research activities.

To tackle this challenge, a web scraper was developed to extract required details from articles and then preprocessing was done using standard normalization techniques. Utilizing this dataset, topic modeling using a BoW as well as using TF-IDF score was performed. Where the latter has demonstrated considerable gains in accuracy ratings (0.34 to 0.483, around 41% improvement to the baseline model score) using the coherence score. Word clouds were also constructed for subject labeling, which could subsequently be utilized for real-time prediction and analysis of a variety of topics as a next proceeding feature.

As a result, this probabilistic model for topic modeling showed more accurate performance and correspondingly high potential. Limitations include:

1. Only one journal is used for evaluating the performance of the proposed model.

2. The proposed model fails to directly model correlation between the occurrences of topics.

3. Human intervention is needed in topic labeling and use of more journals could have shown better results and performance of the proposed model.

Hence, more research and use of the virtuous cycle, in which the findings of data-based prediction techniques are subjected to expert interpretation, the outcomes of which are consequentially employed for data-based prediction methods, is necessary.

The future works include development of a system that can automatically perform trend analysis based on the proposed model in this study. This system will allow users to perform trend analysis.

The fact that there was no involvement of Deep learning, Graph Networks, or similar compounded domains and it still performs well in terms of coherence score when compared to the state of art model, which eventually entailed these kinds of complexities but was still unable to make some good differences in accuracy scores from this proposed model, is the literature's key accomplishment.

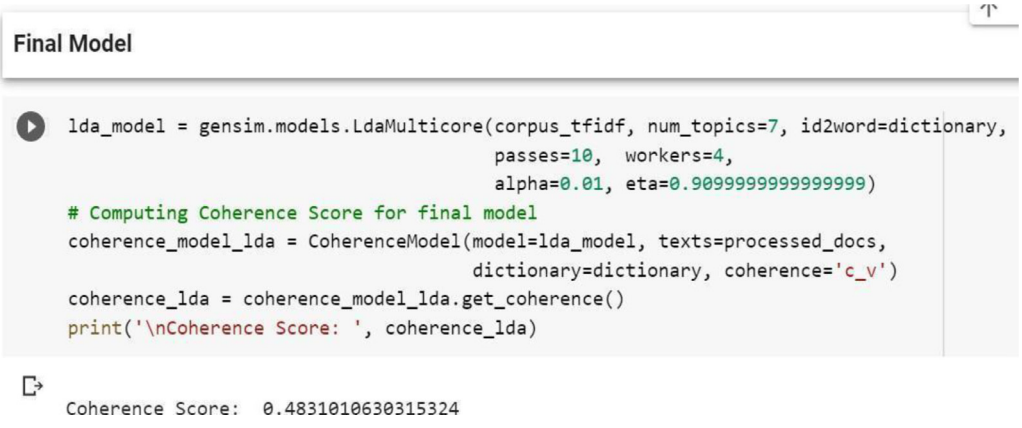

Fig. 9. Execution of Proposed Model.

Fig. 10. Word Clouds.

Testing Model on Unseen Sample Document

```
unseen_document = "Fuzzy logic is based on the observation that people make decisions based on imprecise and n
#Printing the newly formed Bag of Words
bow_vector = dictionary.doc2bow(preprocess(unseen_document))
print('Bag of Words: {}\n'.format(bow_vector));

#calculating TF-IDF scores
unseen_tfidf = tfidf[bow_vector]
print('TF-IDF Scores: {}\n'.format(unseen_tfidf[:2]));

#Finding topics hidden in the unseen document
for index, score in sorted(lda_model[bow_vector], key=lambda tup: -1*tup[1]):
    print("Score: {}\nTopic Number : {}\nTopic: {}".format(score, index, lda_model.print_topic(index, 5)), )
```

```
Bag of Words: [(32, 1), (139, 1), (198, 1), (766, 1), (800, 1), (1312, 1), (1709, 1)]

TF-IDF Scores: [(32, 0.1346959018287869), (139, 0.4291711896367745)]

Score: 0.9915134310722351
Topic Number : 1
Topic: 0.006*"fuzzy" + 0.004*"evidence" + 0.004*"intuitionistic" + 0.004*"measure" + 0.003*"theory"
```

Fig. 11. Finding Latent topic in an unseen Document.

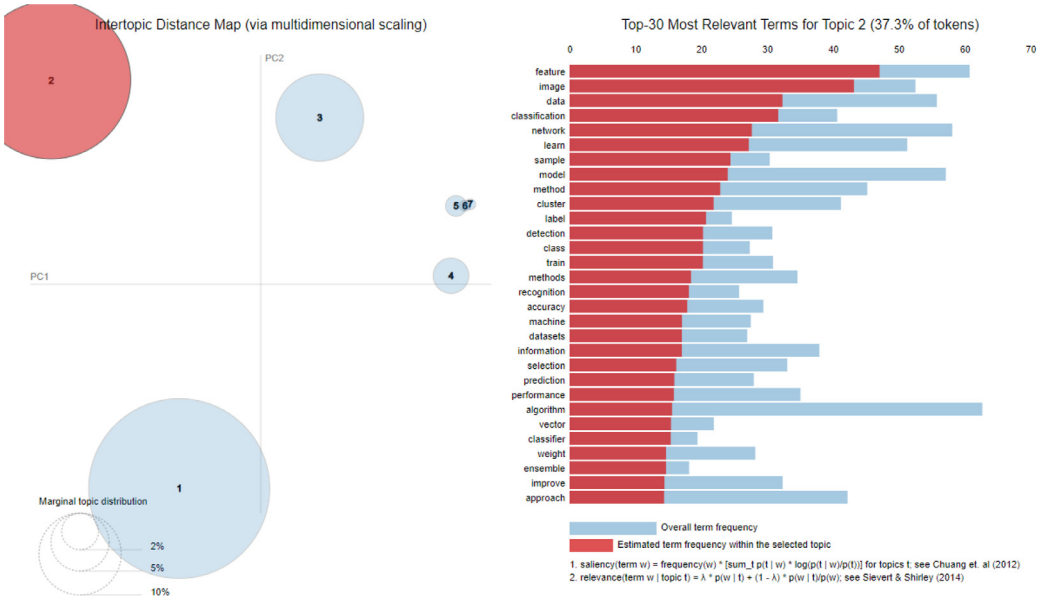

Fig. 12. Visualization of TopicsThe proposed model was also put to the test on unseen documents (shown in Fig. 11). Hence, it's possible to say that an attempt is made to make an unsupervised learning model act like a supervised one. It can be also clearly observed that the topic which is most relevant to the given unseen document is Topic Number 1 (Refer to Word Clouds in Fig. 10) with a score of 0.991. The visualization of topics is depicted in Fig. 12.

Table 3
Dominant Topics.

Doc. No.	Dominant Topic	Topic % contribution	Keywords
0	3.0	0.9920	algorithm, optimization, data, cluster, …
1	3.0	0.9906	algorithm, optimization, data, cluster, …
2	3.0	0.9911	algorithm, optimization, data, cluster, …
3	2.0	0.9901	image, feature, network, detection,
4	4.0	0.9877	fuzzy, knowledge, reason, agents, model,...

Table 4
Comparison of proposed model with the state-of-the-art literature.

Paper Reference	Year of Publication	Accuracy Metric
Proposed Model - LDA utilizing TF-IDF corpus		Coherence Score: 0.483
Sivanandham et al. [28]	2021	Coherence Score: 0.3659
Habibi et al. [33]	2021	Coherence Score: 0.60 (approx.)
Pandur et al. [31]	2020	Semantic Coherence: -100
Albalawi et al. [32]	2020	F-Score: 0.61 (approx.)

Appendix A. Abbreviations used

Abbreviations	Meaning
BoW	Bag of Words
LDA	Latent Dirichlet Allocation
TF-IDF	Term Frequency - Inverse Document Frequency
PMI	Point wise Mutual Information
OR / MS	Operations Research /Management Science
t-SNE	t-Distributed Stochastic Neighbor Embedding
NLP	Natural language Processing

References

[1] Joseph P. Martino, A review of selected recent advances in technological forecasting, Tech. Forecast. Soc. Change 70 (8) (2003) 719–733, doi:10.1016/S0040-1625(02)00375-X.

[2] P. Subramani, P. BD, Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients, Pers. Ubiquit. Comput. (2021) 1–14.

[3] David M. Blei, John D. Lafferty, Topic models, Text Mining (2009) 101–124 Chapman and Hall/CRC, doi:10.1201/9781420059458.

[4] T.N. Nguyen, N.P. Nguyen, C. Savaglio, Y. Zhang, B. Dumba, The role of artificial intelligence (AI) in healthcare data analytics, Int. J. Artif. Intell. Tools 30 (2021) 06 N 08.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[6] K. Yu, L. Lin, M. Alazab, L. Tan, B. Gu, Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system, IEEE Trans. Intell. Transp. Syst. 22 (7) (2020) 4337–4347.

[7] Thomas L. Griffiths, Mark Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5228–5235, doi:10.1073/pnas.0307752101.

[8] B.D. Parameshachari, Big data analytics on weather data: predictive analysis using multi node cluster architecture, Int. J. Comp. Appl. (2022) 0975–8887.

[9] Chaomei. Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, J. Am. Soc. Inform. Sci. Tech. 57 (3) (2006) 359–377, doi:10.1002/asi.20317.

[10] Xuerui Wang, Andrew McCallum, Topics over time: a non-markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, doi:10.1145/1150402.1150450.

[11] M.K. Chowdary, T.N. Nguyen, D.J. Hemanth, Deep learning-based facial emotion recognition for human–computer interaction applications, Neur. Comput. Appl. (2021) 1–18.

[12] Henry. Small, Tracking and predicting growth areas in science, Scientometrics 68 (3) (2006) 595–610, doi:10.1007/s11192-006-0132-y.

[13] Z. Guo, Y. Shen, A.K. Bashir, M. Imran, N. Kumar, D. Zhang, K. Yu, Robust spammer detection using collaborative neural network in Internet-of-Things applications, IEEE Internet of Things J. 8 (12) (2020) 9549–9558.

[14] R.K. Abercrombie, A.W. Udoeyop, B.G. Schlicher, A study of scientometric methods to identify emerging technologies via modeling of milestones, Scientometrics 91 (2012) 327–342, doi:10.1007/s11192-011-0614-4.

[15] B. Rachana, T. Priyanka, K.N. Sahana, T.R. Supritha, B.D. Parameshachari, R. Sunitha, Detection of polycystic ovarian syndrome using follicle recognition technique, Global Trans. Proc. 2 (2) (2021) 304–308.

[16] Murat Bengisu, Ramzi Nekhili, Forecasting emerging technologies with the aid of science and technology databases, Tech. Forecast. Soc. Change 73 (7) (2006) 835–844, doi:10.1016/j.techfore.2005.09.001.

[17] T.D. Ngo, T.T. Bui, T.M. Pham, H.T. Thai, G.L. Nguyen, T.N. Nguyen, Image deconvolution for optical small satellite with deep learning and real-time GPU acceleration, J. Real-Time Image Proc. 18 (5) (2021) 1697–1710.

[18] Nagayoshi Yamashita, Masayuki Numao, Ryutaro Ichise, Predicting research trends identified by research histories via breakthrough research, IEICE Trans. Inf. Syst. 98 (2) (2015) 355–362, doi:10.1587/transinf.2013EDP7435.

[19] L. Tan, K. Yu, F. Ming, X. Chen, G. Srivastava, Secure and resilient artificial intelligence of things: a HoneyNet approach for threat detection and situational awareness, IEEE Consumer Electronics Magazine, 2021.

[20] Nikolaos Aletras, Mark Stevenson, Measuring the similarity between automatically generated topics, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2, 2014 volumeShort Papers, doi:10.3115/v1/E14-4005.

[21] Carson Sievert, Kenneth Shirley, LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the workshop on interactive language learning, visualization, and interfaces, 2014, doi:10.3115/v1/W14-3110.

[22] Christopher J. Gatti, James D. Brooks, Sarah G. Nurre, A historical analysis of the field of OR/MS using topic models, arXiv preprint (2015) arXiv:1510.05154.

[23] Vinodkumar Prabhakaran, et al., Predicting the rise and fall of scientific topics from trends in their rhetorical framing, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1, 2016 VolumeLong Papers, doi:10.18653/v1/P16-1111.

[24] Hyo Choi, Shin, Lee Won Sang, Sohn So Young, Analyzing research trends in personal information privacy using topic modeling, Computers & Security 67 (2017) 244–253, doi:10.1016/j.cose.2017.03.007.

[25] Mario Krenn, Anton Zeilinger, Predicting research trends with semantic and neural networks with an application in quantum physics, Proc. Natl. Acad. Sci. 117 (4) (2020) 1910–1916, doi:10.1073/pnas.1914370116.

[26] T.D. Singh, D. Divyansha, A.V. Singh, A.F.U.R. Khilji, A hybrid classification approach using topic modeling and graph convolutional networks, in: 2020 International Conference on Computational Performance Evaluation (ComPE), 2020, pp. 285–289, doi:10.1109/ComPE49325.2020.9200037.

[27] Kouame Hermann Kouassi, Deshendran Moodley, An analysis of deep neural networks for predicting trends in time series data, in: Southern African Conference for Artificial Intelligence Research, Springer, Cham, 2021, doi:10.1007/978-3-030-66151-9_8.

[28] S. Sivanandham, A. Sathish Kumar, R. Pradeep, R Sridhar, Analysing research trends using topic modelling and trend prediction Reddy V.S., Prasad V.K., Wang J., Reddy K.T.V., Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing, 1325, Springer, Singapore, 2021.

[29] June Young Lee, Sejung Ahn, Dohyun Kim, Deep learning-based prediction of future growth potential of technologies, PLoS One 16 (6) (2021) e0252753, doi:10.1371/journal.pone.0252753.

[30] Sérgio Moro, et al., A text mining and topic modelling perspective of ethnic marketing research, J. Busi. Res. 103 (2019) 275–285, doi:10.1016/j.jbusres.2019.01.053.

[31] Maja Buhin Pandur, Jasminka Dobša, Luka Kronegger, Topic modelling in social sciences-case study of web of science, *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 2020.

[32] Rania Albalawi, Tet Hin Yeap, Morad Benyoucef, Using topic modeling methods for short-text data: A comparative analysis, Front. Artif. Intel. 3 (2020) 42, doi:10.3389/frai.2020.00042.

[33] Muhammad Habibi, et al., Topic modelling of germas related content on Instagram using Latent Dirichlet Allocation (LDA), International Conference on Health and Medical Sciences (AHMS 2020), 2021, doi:10.2991/ahsr.k.210127.060.

[34] Camila Costa Silva, Matthias Galster, Fabian Gilson, Topic modeling in software engineering research, Empir. Softw. Eng. 26 (6) (2021) 1–62, doi:10.5281/zenodo.5280890.