IMPACTS OF VARIOUS TEXT PREPROCESSING METHODS FOR TOPIC MODELING TECHNIQUES

Abstract

Text preprocessing is crucial for topic modeling as it helps to remove noise, correct typos, standardize word usage and apply topic modeling. It involves cleaning and transforming the data to make it suitable for analysis, such as removing stop words, stemming, lemmatization and converting all texts to lowercase. Preprocessing can improve the accuracy of topic modeling by removing noise and irrelevant information. Preprocessed data canimprove the performance of the topic modeling algorithm by providing high-quality input data. In this work, it has been analyzed that the mixture of text data is evaluated with various preprocessing namely, removing stop words, eliminating punctuation, finding the root word using lemmatization, tokenization, creating document term matrix and Latent Dirichlet Allocation. This work can be used to discover hidden topics or themes in a large corpus of text data.

Keywords: Topic modeling, stop words, lemmatization, preprocessing, document term matrix

Authors

S. Alagukumar
Assistant Professor
Department of Computer Applications
Ayya Nadar Janaki Ammal College
Sivakasi
Tamil Nadu, India

R. Lawrance
Director
Department of Computer Applications,
Ayya Nadar Janaki Ammal College,
Sivakasi
Tamil Nadu, India.

I. INTRODUCTION

Topic modeling is a technique used in machine learning and natural language processing (NLP) to discover hidden topics or themes in a large corpus of text data[1]. It can be used to identify patterns and trends in language use, understand the structure of a dataset, and identify topics that are not readily apparent. In recent years machine learning and natural language processing algorithms have been used to analyze the massive amount of text data available online, including topic modeling techniques [2]. The data have to be cleaned before the analysis, the preprocessing is tokenizing, standardizing, cleaning, removing stop words, and stemming. This text preprocessing is highly influenced by the decision-making for the text analysis [3].There are common algorithms used in topic modeling including Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NM).However, preprocessing is an important step in topic modeling as it helps in improving the quality and accuracy of the results. It involves cleaning and transforming the raw text data to make it suitable for analysis. Preprocessing methods helpto reducethe noise and irrelevant information to allow topic modeling algorithms to focus on extracting meaningful patterns and topics from the data. Section 2, related works are presented. In section 3, Methods are explained. Section 4, describes experimental results and discussion.

II. RELATED WORKS

The most recent innovation in text mining is topic modeling. It is a statistical approach for illuminating the underlying semantic structure in huge document collections [4]. Vayansky and kumar[5] have stated that A well-liked analytical tool for assessing data is topic modeling. Instead, a lot of academics use latent dirichlet analysis, which, although adaptable and versatile. Churchill and Singh [6] have mentioned that to find the most common topics in a text corpus, topic models have been used for literature, newspapers, and social media posts. Chen *et al*. [7] stated that learning analytics (LA) has become an increasingly active field focusing on leveraging learning process data to understand and improve teaching and learning using structural topic modeling. Gencoglu *et al*.[8] have suggested that without the assistance of automated text analysis methods, the large-scale analysis of written text responses is extremely time-consuming. Thus, the automated methods are valuable in assisting in the analysis of students' written text responses. Shin *et al.* [9] have analyzed the teaching of mathematics to students with disabilities using word networks and topic modeling, which is effective in extracting meaningful categories and themes from large datasets. Abdelrazek *et al*.[10] have topic modeling that is used in information retrieval to infer the hidden themes in a collection of documents and thus provides an automatic means to organize, understand and summarize large collections of textual information. Topic models also offer an interpretable representation of documents used in several downstream Natural Language Processing (NLP) tasks. Modeling techniques vary from probabilistic graphical models to the more recent neural models.

III. METHODOLOGY

The proposed method has six phases, namely, i. removing stop words, ii. removing punctuations, iii. lemmatization, iv. tokenization, v. document term matrix, and vi. analysis of topics using Latent Dirichlet Allocation. The flow of the proposed methodology is explained in the procedure.

1. **Procedure:**

   Step 1: Read the multiple documents
   Step 2: Identify and Remove the stopwords
   Step 3: Identify and Remove the punctuation
   Step 4: Identify the root word using lemmatization
   Step 5: Break the text features into tokens
   Step 6: Prepare the document term matrix of preprocessed data
   Step 7: Build the model using the LDA algorithm
   Step 8: Identify the hidden information

2. **Removing Stop Words:** The first step of preprocessing is identifying the stop words. Stopwords [11] are high-frequency words that do not carry much meaning in a sentence or document, such as "the", "a", "and", etc. In Natural Language Processing (NLP), stopwords are removed or filtered out to reduce noise and improve text analysis, as they often dominate bag-of-words representations. By removing stopwords, topic modeling can focus on more meaningful words and gain insights into the topics, sentiments, and intentions of the text. Removing stop words can help improve the accuracy of natural language processing tasks such as text classification, sentiment analysis, and information retrieval.

3. **Removing Punctuation:** The important preprocessing is identifying punctuation from the multiple documents. Removing punctuation [11] can help improve the accuracy of topic modeling by reducing the impact of stop words and special characters, which can be noisy and distract from the underlying topics. By removing punctuation, the model can focus more on the meaning of the words and their relationships.

4. **Lemmatization:** The most important preprocessing is identifying root words from multiple documents. Lemmatization is a natural language processing technique that reduces words to their base or root form, called a lemma. It is commonly used in topic modeling to normalize and unify words with similar meanings. Performing lemmatization on text data before applying topic modeling algorithms, It can improve the accuracy and interpretability of the resulting topics. Lemmatization considers factors such as verb tense, plural/singular forms, and word endings to transform words into their canonical form. This helps in reducing noise and redundancy in the data, which is important for obtaining meaningful and coherent topics in topic modeling [12].

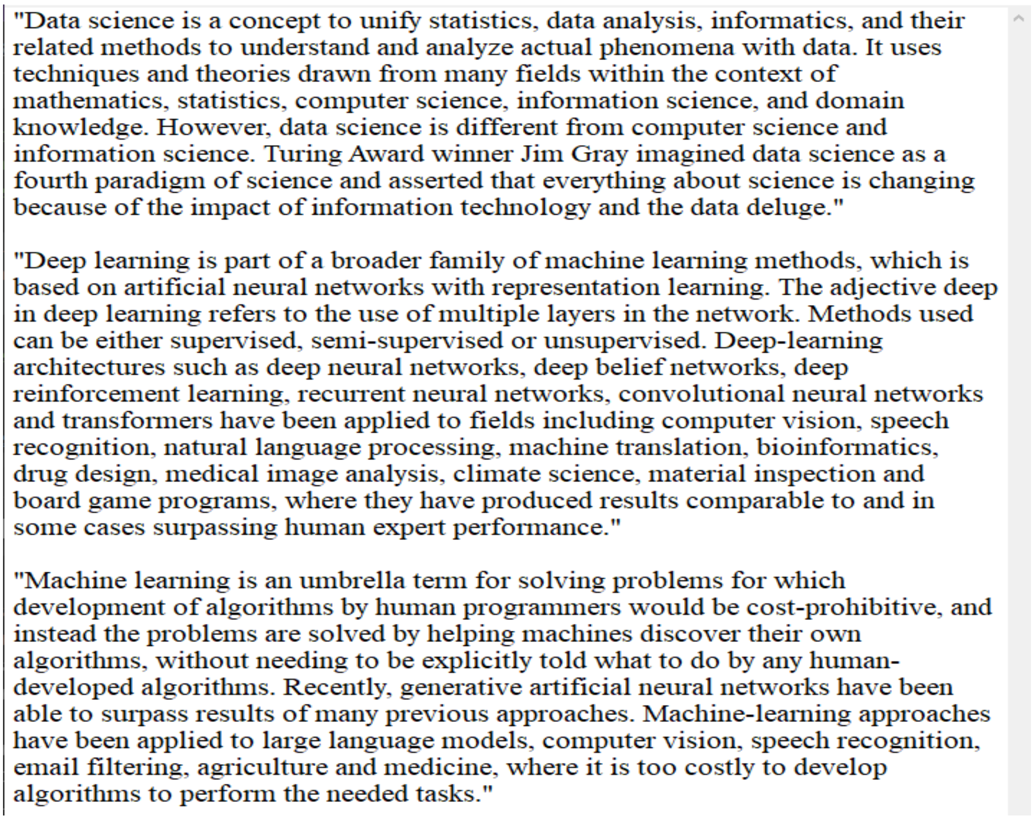
5. **Tokenization:** One of the most vital preprocessing is tokenization to create a document term matrix and word vector matrix. Tokenization [13] is the process of breaking down text into individual units called tokens. In the context of topic modeling, tokenization is crucial as it is the first step in preparing the text data for analysis. In topic modeling, tokens are often words or phrases that represent meaningful units of information. By tokenizing the text, you can create a bag of words or a sequence of tokens that can be used as input for topic modeling algorithms. There are different ways to tokenize text such as using whitespace or punctuation as delimiters. Tokenization is an important technique to extract valuable insights and discover hidden patterns within textual data.

6. **Document term matrix:** A document term matrix [14] is a way to represent a collection of text documents in a numerical format that can be easily processed by machine learning algorithms. In this matrix, rows represent documents and columns represent terms or words in the documents. Each entry in the matrix represents the frequency of a particular term in a particular document. This matrix is useful for text mining tasks such as sentiment analysis, topic modeling, and document clustering. It allows for efficient analysis and comparison of large text datasets.

7. **Latent Dirichlet Allocation:** Latent Dirichlet Allocation is a popular unsupervised topic modeling algorithm [15] used to discover hidden topics in a corpus. It works by representing documents as mixtures of topics, where each topic is characterized by a distribution over words. LDA is widely used in NLP applications, including text classification, document clustering, and information. In this work, the LDA algorithm is used to analyze multiple documents to find the hidden topics.

IV. RESULTS AND DISCUSSION

In this work, the data set is created using three different text documents such as data science machine learning and deep learning. The work is carried out by usingNatural Language Tool Kit (nltk) and gensim package of python. Figure 1 represents the raw data with contains stop words, and punctuations. The stop words are identified and eliminated in the raw data which are shown in figure 2. Then the punctuations areidentified and eliminated from the stop words, which are shown in Figure 3. Then, the data are reduced to their base or root form using lemmatization, which is shown in Figure 4. Then the data are broken down into a sequence of text into smaller units, or tokens. Figure 5 represents the tokenized data. Then the tokenized data are converted into the document term matrix. Document term matrix that stores the frequency of term appearances in a document, where the rows represent the terms and the columns represent the documents. Figure 6 represents the document term matrix. Finally, the term matrix is passed into the LDA algorithm to discover hidden topics in a corpus of text documents. It's a generative model that assumes each document is a mixture of topics, and each topic is a mixture of words. Figure 7 represents the clustered terms of different topics.

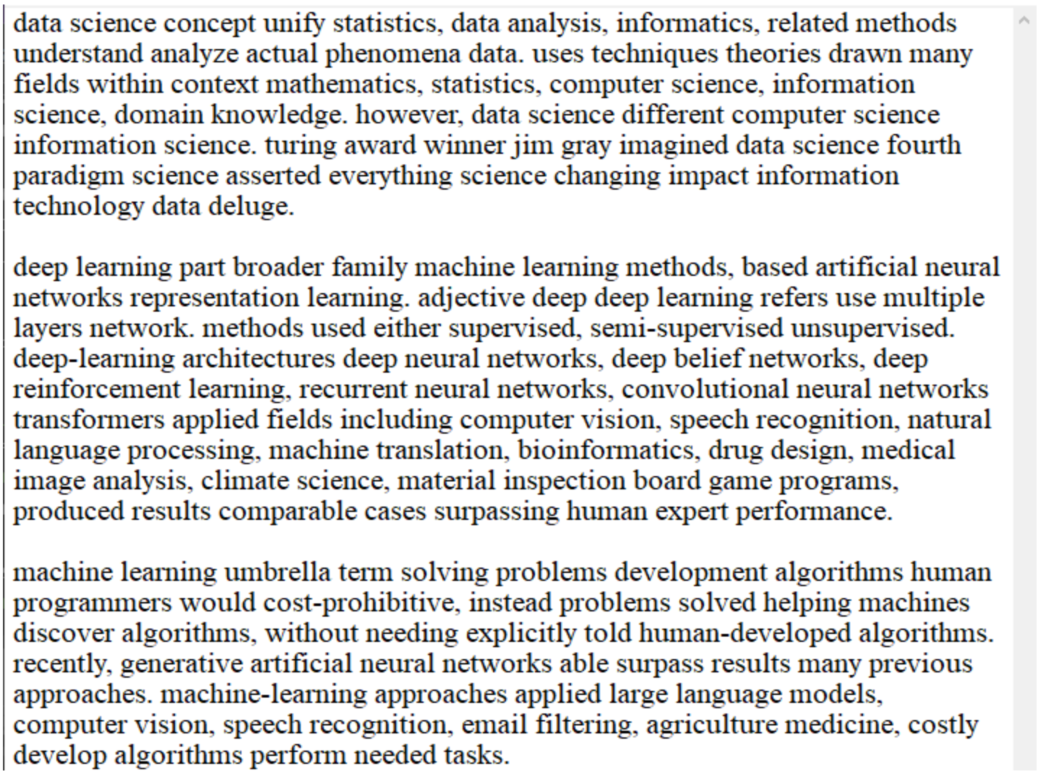


"Data science is a concept to unify statistics, data analysis, informatics, and their related methods to understand and analyze actual phenomena with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a fourth paradigm of science and asserted that everything about science is changing because of the impact of information technology and the data deluge."

"Deep learning is part of a broader family of machine learning methods, which is based on artificial neural networks with representation learning. The adjective deep in deep learning refers to the use of multiple layers in the network. Methods used can be either supervised, semi-supervised or unsupervised. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, convolutional neural networks and transformers have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance."

"Machine learning is an umbrella term for solving problems for which development of algorithms by human programmers would be cost-prohibitive, and instead the problems are solved by helping machines discover their own algorithms, without needing to be explicitly told what to do by any human-developed algorithms. Recently, generative artificial neural networks have been able to surpass results of many previous approaches. Machine-learning approaches have been applied to large language models, computer vision, speech recognition, email filtering, agriculture and medicine, where it is too costly to develop algorithms to perform the needed tasks."

Figure 1: Raw Data

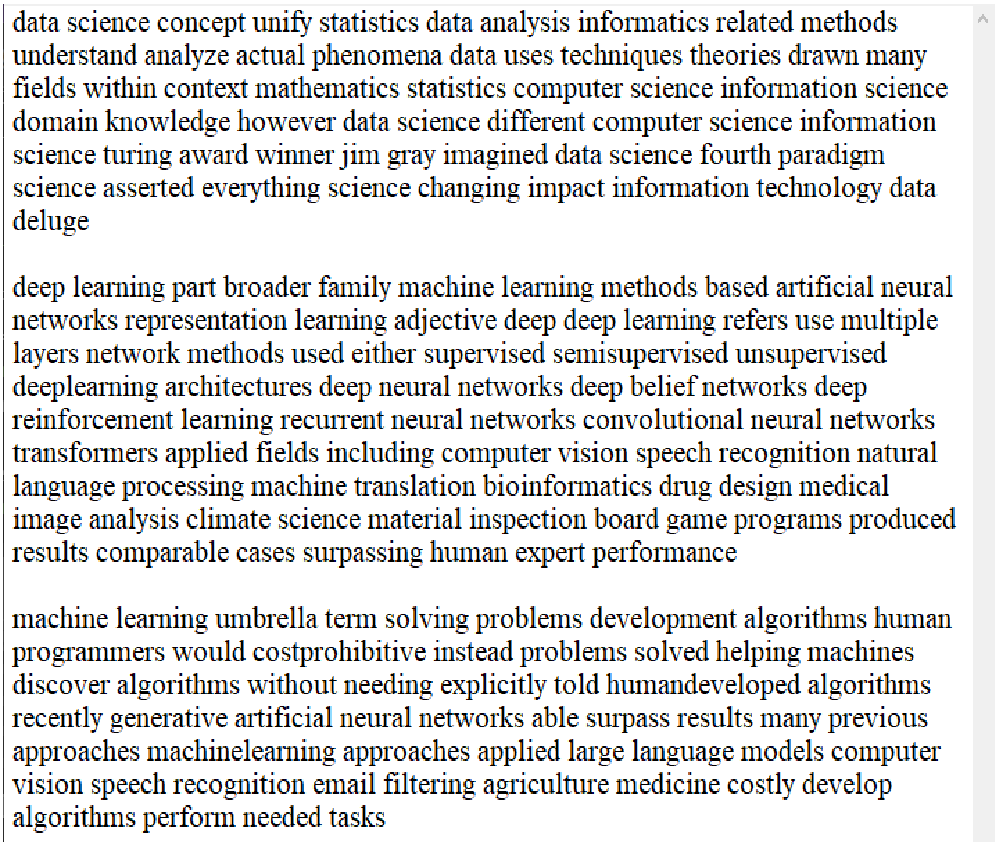


data science concept unify statistics, data analysis, informatics, related methods understand analyze actual phenomena data. uses techniques theories drawn many fields within context mathematics, statistics, computer science, information science, domain knowledge. however, data science different computer science information science. turing award winner jim gray imagined data science fourth paradigm science asserted everything science changing impact information technology data deluge.

deep learning part broader family machine learning methods, based artificial neural networks representation learning. adjective deep deep learning refers use multiple layers network. methods used either supervised, semi-supervised unsupervised. deep-learning architectures deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, convolutional neural networks transformers applied fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection board game programs, produced results comparable cases surpassing human expert performance.

machine learning umbrella term solving problems development algorithms human programmers would cost-prohibitive, instead problems solved helping machines discover algorithms, without needing explicitly told human-developed algorithms. recently, generative artificial neural networks able surpass results many previous approaches. machine-learning approaches applied large language models, computer vision, speech recognition, email filtering, agriculture medicine, costly develop algorithms perform needed tasks.

Figure 2: Stop Words Removed Data

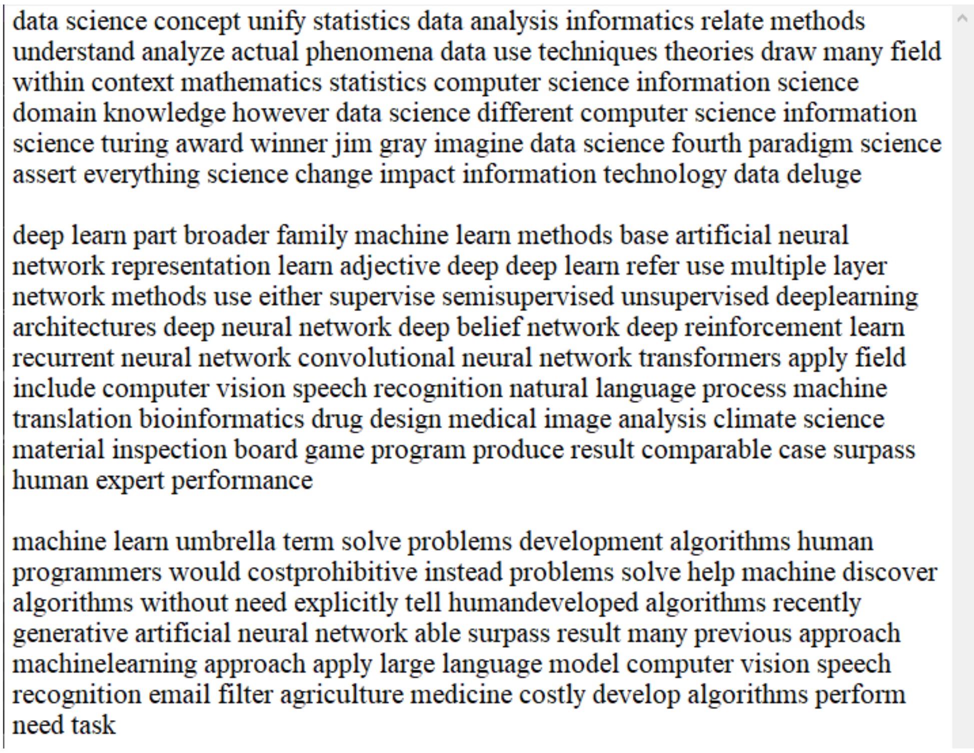


data science concept unify statistics data analysis informatics related methods understand analyze actual phenomena data uses techniques theories drawn many fields within context mathematics statistics computer science information science domain knowledge however data science different computer science information science turing award winner jim gray imagined data science fourth paradigm science asserted everything science changing impact information technology data deluge

deep learning part broader family machine learning methods based artificial neural networks representation learning adjective deep deep learning refers use multiple layers network methods used either supervised semisupervised unsupervised deeplearning architectures deep neural networks deep belief networks deep reinforcement learning recurrent neural networks convolutional neural networks transformers applied fields including computer vision speech recognition natural language processing machine translation bioinformatics drug design medical image analysis climate science material inspection board game programs produced results comparable cases surpassing human expert performance

machine learning umbrella term solving problems development algorithms human programmers would costprohibitive instead problems solved helping machines discover algorithms without needing explicitly told humandeveloped algorithms recently generative artificial neural networks able surpass results many previous approaches machinelearning approaches applied large language models computer vision speech recognition email filtering agriculture medicine costly develop algorithms perform needed tasks

Figure 3: Punctuation Removed Data

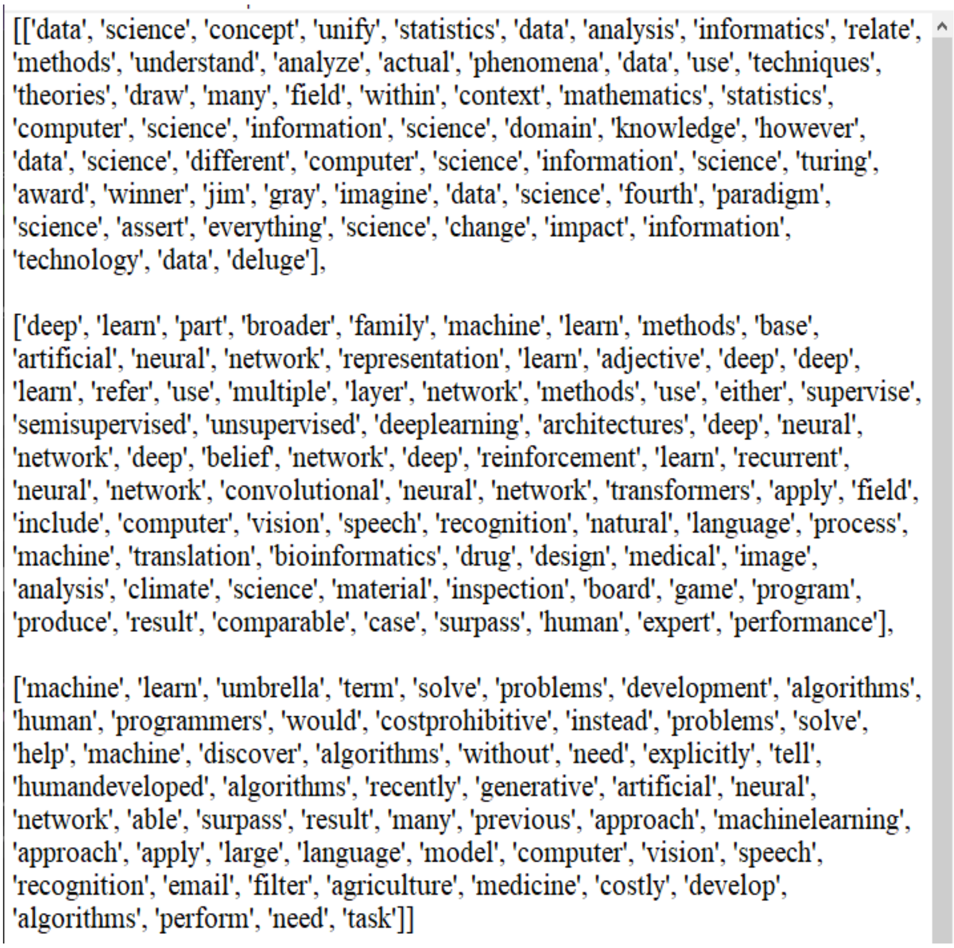


data science concept unify statistics data analysis informatics relate methods understand analyze actual phenomena data use techniques theories draw many field within context mathematics statistics computer science information science domain knowledge however data science different computer science information science turing award winner jim gray imagine data science fourth paradigm science assert everything science change impact information technology data deluge

deep learn part broader family machine learn methods base artificial neural network representation learn adjective deep deep learn refer use multiple layer network methods use either supervise semisupervised unsupervised deeplearning architectures deep neural network deep belief network deep reinforcement learn recurrent neural network convolutional neural network transformers apply field include computer vision speech recognition natural language process machine translation bioinformatics drug design medical image analysis climate science material inspection board game program produce result comparable case surpass human expert performance

machine learn umbrella term solve problems development algorithms human programmers would costprohibitive instead problems solve help machine discover algorithms without need explicitly tell humandeveloped algorithms recently generative artificial neural network able surpass result many previous approach machinelearning approach apply large language model computer vision speech recognition email filter agriculture medicine costly develop algorithms perform need task

Figure 4: Lemmatized Data

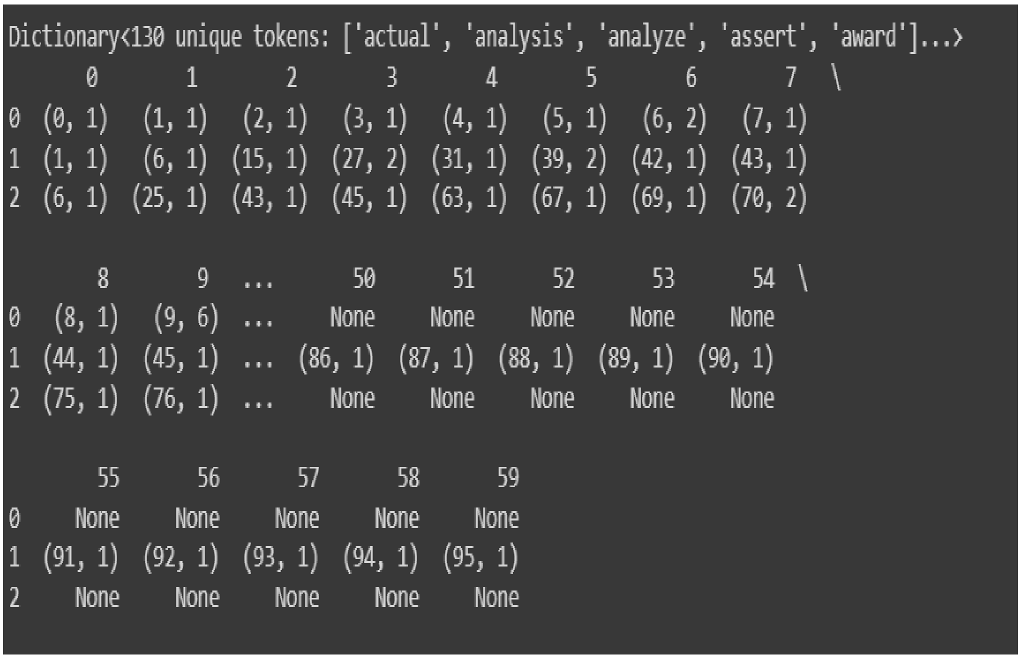


```
[['data', 'science', 'concept', 'unify', 'statistics', 'data', 'analysis', 'informatics', 'relate',
'methods', 'understand', 'analyze', 'actual', 'phenomena', 'data', 'use', 'techniques',
'theories', 'draw', 'many', 'field', 'within', 'context', 'mathematics', 'statistics',
'computer', 'science', 'information', 'science', 'domain', 'knowledge', 'however',
'data', 'science', 'different', 'computer', 'science', 'information', 'science', 'turing',
'award', 'winner', 'jim', 'gray', 'imagine', 'data', 'science', 'fourth', 'paradigm',
'science', 'assert', 'everything', 'science', 'change', 'impact', 'information',
'technology', 'data', 'deluge'],

['deep', 'learn', 'part', 'broader', 'family', 'machine', 'learn', 'methods', 'base',
'artificial', 'neural', 'network', 'representation', 'learn', 'adjective', 'deep', 'deep',
'learn', 'refer', 'use', 'multiple', 'layer', 'network', 'methods', 'use', 'either', 'supervise',
'semisupervised', 'unsupervised', 'deeplearning', 'architectures', 'deep', 'neural',
'network', 'deep', 'belief', 'network', 'deep', 'reinforcement', 'learn', 'recurrent',
'neural', 'network', 'convolutional', 'neural', 'network', 'transformers', 'apply', 'field',
'include', 'computer', 'vision', 'speech', 'recognition', 'natural', 'language', 'process',
'machine', 'translation', 'bioinformatics', 'drug', 'design', 'medical', 'image',
'analysis', 'climate', 'science', 'material', 'inspection', 'board', 'game', 'program',
'produce', 'result', 'comparable', 'case', 'surpass', 'human', 'expert', 'performance'],

['machine', 'learn', 'umbrella', 'term', 'solve', 'problems', 'development', 'algorithms',
'human', 'programmers', 'would', 'costprohibitive', 'instead', 'problems', 'solve',
'help', 'machine', 'discover', 'algorithms', 'without', 'need', 'explicitly', 'tell',
'humandeveloped', 'algorithms', 'recently', 'generative', 'artificial', 'neural',
'network', 'able', 'surpass', 'result', 'many', 'previous', 'approach', 'machinelearning',
'approach', 'apply', 'large', 'language', 'model', 'computer', 'vision', 'speech',
'recognition', 'email', 'filter', 'agriculture', 'medicine', 'costly', 'develop',
'algorithms', 'perform', 'need', 'task']]
```

Figure 5: Tokenized Data

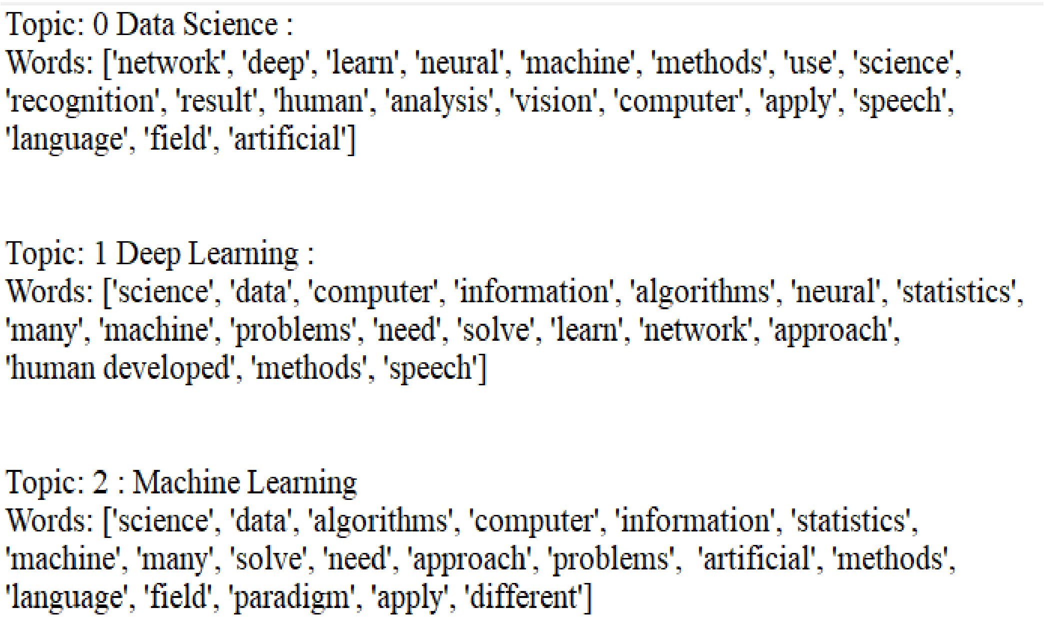


```
Dictionary<130 unique tokens: ['actual', 'analysis', 'analyze', 'assert', 'award']...>
        0       1       2       3       4       5       6       7  \
0  (0, 1)  (1, 1)  (2, 1)  (3, 1)  (4, 1)  (5, 1)  (6, 2)  (7, 1)
1  (1, 1)  (6, 1) (15, 1) (27, 2) (31, 1) (39, 2) (42, 1) (43, 1)
2  (6, 1) (25, 1) (43, 1) (45, 1) (63, 1) (67, 1) (69, 1) (70, 2)

         8        9  ...       50       51       52       53       54  \
0   (8, 1)   (9, 6)  ...     None     None     None     None     None
1  (44, 1)  (45, 1)  ...  (86, 1)  (87, 1)  (88, 1)  (89, 1)  (90, 1)
2  (75, 1)  (76, 1)  ...     None     None     None     None     None

        55       56       57       58       59
0     None     None     None     None     None
1  (91, 1)  (92, 1)  (93, 1)  (94, 1)  (95, 1)
2     None     None     None     None     None
```

Figure 6: Document Term Matrix



```
Topic: 0 Data Science :
Words: ['network', 'deep', 'learn', 'neural', 'machine', 'methods', 'use', 'science',
'recognition', 'result', 'human', 'analysis', 'vision', 'computer', 'apply', 'speech',
'language', 'field', 'artificial']


Topic: 1 Deep Learning :
Words: ['science', 'data', 'computer', 'information', 'algorithms', 'neural', 'statistics',
'many', 'machine', 'problems', 'need', 'solve', 'learn', 'network', 'approach',
'human developed', 'methods', 'speech']


Topic: 2 : Machine Learning
Words: ['science', 'data', 'algorithms', 'computer', 'information', 'statistics',
'machine', 'many', 'solve', 'need', 'approach', 'problems',  'artificial', 'methods',
'language', 'field', 'paradigm', 'apply', 'different']
```

Figure 7: Clustered Topics

V. CONCLUSION

In this paper, it has been proposed a method to analyse the mixture of text data using various pre-processing technique and topic modeling. Preprocessing is a crucial step before applying topic modeling. The topic model is to focus on the meaningful content and identify meaningful topics from a mixture of datasets. In this work, it has been analyzed mixture of documents and evaluated with various preprocessing namely, removing stop words, eliminating punctuation, finding the root word using lemmatization, tokenization, creating document term matrix and Latent Dirichlet Allocation. From this work, it has been found that by properly preprocessing the data, the accuracy of the topic modeling results also improved. This work helps organizations to gain a deeper understanding of hidden information from various topics.

REFERENCES

[1] Barde, B. V., andBainwad, A. M. (2017). An overview of topic modeling methods and tools. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 745-750). IEEE.

[2] Albalawi, R., Yeap, T. H., and Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. Frontiers in artificial intelligence, 3, 42.

[3] Anandarajan, M., Hill, C., Nolan, T., Anandarajan, M., Hill, C., and Nolan, T. (2019). Text preprocessing. Practical text analytics: Maximizing the value of text data, 45-59.

[4] Kherwa, P., and Bansal, P. (2019). Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems, 7(24).

[5] Vayansky, I., and Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

[6] Churchill, R., and Singh, L. (2022). The evolution of topic modeling. ACM Computing Surveys, 54(10s), 1-35.

[7] Chen, X., Zou, D., andXie, H. (2022). A decade of learning analytics: Structural topic modeling based bibliometric analysis. Education and Information Technologies, 27(8), 10517-10561.

[8] Gencoglu, B., Helms-Lorenz, M., Maulana, R., Jansen, E. P., andGencoglu, O. (2023). Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data. Computers & Education, 193, 104682.

[9] Shin, M., Ok, M. W., Choo, S., Hossain, G., Bryant, D. P., and Kang, E. (2023). A content analysis of research on technology use for teaching mathematics to students with disabilities: word networks and topic modeling. International Journal of STEM Education, 10(1), 1-23.

[10] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2022). Topic modeling algorithms and applications: A survey. Information Systems, 102131.

[11] Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

[12] Boban, I., Doko, A., andGotovac, S. (2020). Sentence retrieval using stemming and lemmatization with different length of the queries. Advances in Science, Technology and Engineering Systems, 5(3), 349-354.

[13] Rahutomo, R., Lubis, F., Muljo, H. H., andPardamean, B. (2019, August). Preprocessing methods and tools in modelling japanese for text classification. In 2019 International Conference on Information Management and Technology (ICIMTech) (Vol. 1, pp. 472-476). IEEE.

[14] Anandarajan, M., Hill, C., Nolan, T., Anandarajan, M., Hill, C., & Nolan, T. (2019). Term-document representation. Practical Text Analytics: Maximizing the Value of Text Data, 61-73.

[15] Momtazi, S., andNaumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(5), 346-353.