International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

From Data to Insight: Topic Modeling and Automatic Labeling Strategies

Rana F. Najeeb [1], **Ban N. Dhannoon** [2] **and Farah Qais Alkhalidi** [3]

[1] *Computer Science of Mustansiriyah University, Baghdad, Iraq*
[2] *Computer Science of Al-Nahrain University, Baghdad, Iraq*
[3] *Computer Science of Mustansiriyah University, Baghdad, Iraq*

E-mail address: rana19.najeeb@uomustansiriyah.edu.iq, ban.n.dhannoon@nahrainuniv.edu.iq , farahqaa@uomustansiriyah.edu.iq

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: Researchers usually present and synthesize their findings in scientific publications. For this reason, it is essential to analyze their substance to understand a subject. This study suggests improving the topic modeling in a collection of conference papers on Neural Information Processing Systems (NIPS) released between 1987 and 2017. Two goals of this study were achieved: producing more coherent topics and topic automatic labeling. The first goal was achieved through five phases, text pre-processing phase, reduction phase using a new method called RS-LW (Reduced Sentences Based on Length and Weight), which removes the sentences of shorter length, then calculates the weight for the remaining sentences and removes approximately 25% of the less weight sentences. Sentence embedding phase using S-BERT (Sentence-Bidirectional Encoder Representation from Transformer), Reducing the dimensionality of the sentences embedding phase by utilizing UMAP (Uniform Manifold Approximation and Projection). Lastly, the use of HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to organize comparable documents. The experimental findings demonstrate that the use of the proposed RS-LW phase has produced more cohesive topics. This has led to improvements in topic coherence by (0.593), and topic diversity performance by (0.96). Though topic modeling extracts the most salient sentences describing latent topics from text collections, an appropriate label has not yet been identified. The second goal was achieved by suggesting a new method to generate the keywords by accessing the authors profile in Google Scholar and extracting the interests for use in automatically labeling the topics.

Keywords: Deep Learning, Topic Modelling, Automatic Topic Labeling, S-BERT, Pre-trained Language Model.

1. INTRODUCTION

Comprehend the effectiveness and shortcomings of its creators and owners. These data can be, by definition, social media, scientific, biological, or operational, demonstrating the diversity of various datasets [1, 2]. News headlines, tweets, social media posts, blog entries, user comments, news stories, scientific articles, and other sources are some of the many sources that provide textual information [3].

Effective machine learning techniques and algorithmic models are necessary for correct data interpretability in order to achieve the goals [4].

The detection of inherent topics and semantic structure in a large-scale text collection has drawn the attention of statisticians, analysts, and academicians among the myriad approaches, theories, and applications in the field of text mining, including document clustering, text classification, extraction of information, named entity recognition, text analytics, and so forth [5].

The other popular method of this type is topic modeling, which discovers the underlying hidden latent semantic representations in a given collection of documents [6]. Topic models analyze text documents to discover underlying themes (referred to as topics) they contain, and how those themes (topics) are connected [7]. In topic models, a topic is a collection of the most probable words in the cluster.

Topic modeling approaches are a form of unsupervised machine learning techniques since the topics, and mixture

E-mail:author's email

parameters are not known and are inferred solely from the data. In other words, it is not trained on already tagged or labeled data. The most commonly used probabilistic topic modeling technique is LDA [8,9], and another very foundational topic modeling technique is pLSA [10,11].

Both these models are extensively used for topic modeling and have been modified and extended for many new models. In 2018 Google developed the generative transformer BERT model based on neural networks [12,13].

The pre-trained BERT model generates contextual embedding for every single word or token in the text. Using BERT embedding in the process of topic modeling can be considered a kind of transfer learning. The BERT model is used for obtaining sentence embedding as an averaged embedding overall of its constituent words (tokens) [14,15].

Instead of BERT, other similar models to generate context-dependent embedding for sentences can be used [16,17]. Automatic topic labeling is another essential aspect and subtask of topic modeling [18].

Automatic topic labeling is an algorithmic process of generating/selecting phrases or sentences that describe a topic in the best form. This field of topic labeling is poorly studied and developed [18].

One of the earliest techniques for labeling text was the hand labeling of topics. The manual labeling of themes was a key component of certain previous labeling strategies.

When labeling tasks manually, a skilled user selects a label that best encapsulates the topics by taking a close look at a set of words that are related to a particular issue. This manual method is labor-intensive and slow since it involves several human interactions [19].

The contribution of this study is described as follows:

- ❖ Propose a new RS-LW approach which improves the extracting of topics that represent the collection of documents or corpus through topic modeling process.

- ❖ proposed a new method for analysis and labeling the generated topics automatically.

There are five sections in this study; Section 1 introduces the topic modeling and labeling problem with its importance. Section 2 examines related researches on different topic modeling and labeling methods. Section 3 describing the dataset used. The proposed topic modeling and labeling methods and various steps are discussed in section 4 and 5. Further, Section 6 gives an evaluation, Section 7 discusses the results obtained and the conclusion that are given in Section 8.

2. RELATED WORKS

The Topic Modeling methods efficiently extract themes, hotspots, and current trends from massive text corpora through processing. Users may find it easier to grasp newly discovered topics when these words are meaningfully labeled in each topic.

Topic labeling is to automatically generate semantically appropriate labels for the text categorization or word group. This review of the literature provides an in-depth examination of studies that deal with the problem of extracting labels and topics from corpus or text collections [9, 20].

A. Topic Modeling

The specifics of the proposed approach involve assessing sentence probabilities within the text corpus and clustering sentence embedding.

The thesis in [7] delineates a method for generating semantically meaningful topics by leveraging contextual embedding like BERT and Sentence-BERT.

In [14], the authors develop a topic clustering approach based on BERT-LDA joint embedding that takes contextual semantics and topic narrative into account. They use cluster text embedding using the HDBSCAN algorithm and a class-based TF-IDF (c-TF-IDF) technique to construct topic representations.

The study in reference [15] centers on the application of Transformer models, such SBERT, to topic modeling and evaluates the degree to which these models reveal significant structure. Describe what was learned during the COVID-19 pandemic and the key advantages of using BERTopic in large-scale data analysis.

The text vectorization technique presented by the authors in [21] combines transfer learning with a topic model. First, in order to model the data from the text and extract its keywords, the topic model is chosen in order to identify the primary information included in the data. Next, model transfer learning is performed to produce vectors that are used in the computation of text similarity between texts, with the use of the pre-trained model's (The BERT algorithm) model.

In [22], a study that combines the benefits of both BERT and LDA topic modeling techniques presents a unified clustering-based framework that mines significant themes from massive text corpora.

The paper in [23] describes how the BERTopic architecture uses K-means clustering and Kernel Principal Component Analysis (Kernel PCA).

The primary goal of [24] was to identify word topic clusters in pre-trained language models using the BERT and DistilBERT. It was discovered that the attention framework is crucial to the modeling of this kind of word topic clustering.

In [25] the authors proposed an LDA-Bert public opinion topic mining model to solve the problem that LDA ignores the context semantics, and the topic distribution is biased towards high-frequency words.

The most important of this study [26] is to identify the types of information and their effective impact an analytical theory for analyzing fake news related to the Coronavirus (Covid-19). It combines the idea of sentiment analysis (SA) and Topic Model (TM) for source optimization of a large amount of unstructured material Texts by looking at the feelings of words. The dataset contains 10,254 custom addresses from all over the world The essential and required elements were collected, and the documents were applied to SA to name the next dataset Tags. Among the TM models evaluated, hidden Dirichlet allocation (LDA) shows the highest satisfactory level 0.66 for 20 products leading to affect emotions and 0.573 for 18 false positives leading to Subjects excelled in non-defensive matrix factorization (NMF) (significant value: 0.43) and latent semantic analysis. (LSA) (Significance ratio: 0.40).

B. Topic Labeling

Under some guidance, [27] recommends providing a topic with a succinct label that captures its principal concept or topic. Using Wikipedia article titles as label possibilities, neural embedding for words and documents was computed to choose the best labels for the subjects. The authors in [28] introduce a fourth topic labeler that extracts representative sentences, using Dirichlet smoothing to add contextual information. This sentence-based labeler provides strong surrogate candidates when n-gram topic labelers fall short of providing relevant labels, leading up to 94% topic covering.

The study in [29] presented a novel two-phase neural embedding system that incorporates a graph-based ranking method that is mindful of redundancy.

It illustrated how topic names, sentence presentations, and automatic topic labeling tasks could benefit from the application of pre-trained neural embedding.

In [30] The paper proposes a method to generate labels automatically to represent each topic based on a labeling strategy to filter candidate labels and then apply sequence-to-sequence labelers. The objective of the method is to get a meaningful label for the result of the Latent Dirichlet Allocation algorithm.

The article in [31] proposes an automatic tagging model that includes BERT and word2vec. The model has been validated There is data for electrical tools.

Within the model, PERT's method works to obtain Shallow text marks. Moreover, lightweight text optimization is used to solve the diversity problem They are cut off when there are a number of suitable stickers. Finally, the word2vec model was used Deep text analysis.

3. Dataset Description

An essential component of the NIPS conference, the NIPS (Conference on Neural Information Processing Systems) stimulates AI research in fields such as computer vision and natural language processing (NLP).

A dataset that spans 30 years—from 1987 to 2016—contains 7280 publications. It is well-liked by its contributors and accessible to the general public on Kaggle [30].

The following characteristics (see dataset on the link NIPS Papers | Kaggle (NIPS Papers | Kaggle)) with size file (408 MB) which characterize each paper by : id, year of publication, title, PDF name, event type, abstract, and full text. Units.

4. METHODOLGY

The proposed study analyzes conference papers from Neural Information Processing Systems (NIPS). It utilizes Sentence-BERT, UMAP, and HDBSCAN, presenting a comprehensive workflow for in-depth topic modeling as shown in Figure 1. It compares established methods and introduces a new approach, RS-LW, for efficient analysis of extensive text datasets.

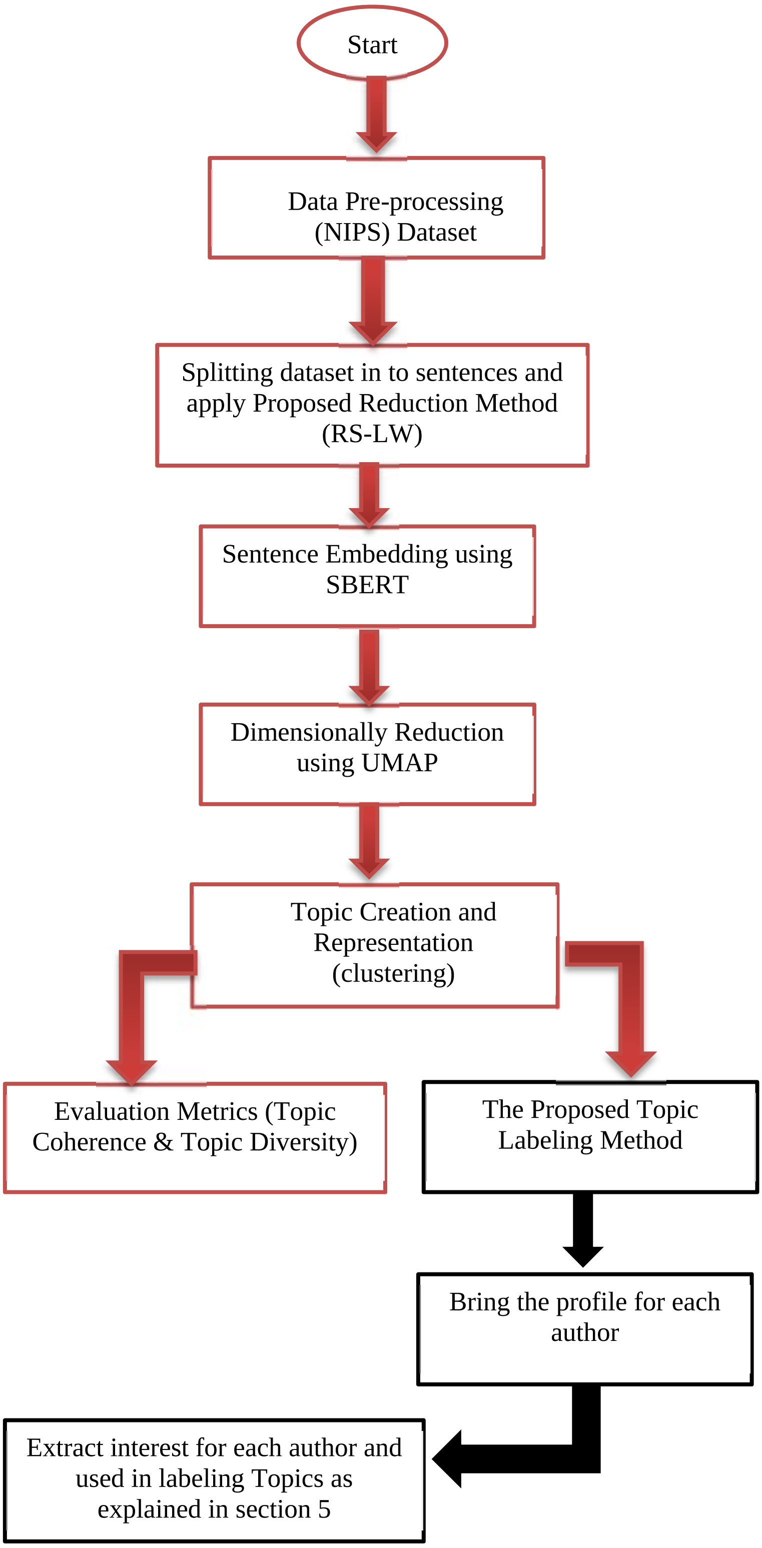
Start
Data Pre-processing (NIPS) Dataset
Splitting dataset in to sentences and apply Proposed Reduction Method (RS-LW)
Sentence Embedding using SBERT
Dimensionally Reduction using UMAP
Topic Creation and Representation (clustering)
Evaluation Metrics (Topic Coherence & Topic Diversity)
The Proposed Topic Labeling Method
Bring the profile for each author
Extract interest for each author and used in labeling Topics as explained in section 5

Figure 1: Flowchart of the Methodology Work Steps

A. Data Pre-processing

Since text data can have different formats and contain errors, data cleaning is necessary. The following steps are taken to achieve a targeted text format for each paper in the dataset:

- Removing the text between Title and Abstract.
- Removing acknowledgments, and references from all paper text in the list of text, as they do not contribute added value to the evaluation.

- Using stop words in the Natural Language Toolkit (NLTK) library and add a list of customized stop words filtered from the dataset and convert all to lowercase.
- Create two lists including (a list of titles and a list of texts).
- Removing stop words and words less than 4 letters from two lists (text, title).
- Convert the two lists (littles and texts) to lowercase and remove new lines from the list of text to make all the text in one line.
- Removing special characters and digits (except dot) from the two lists (titles and texts).
- Replace multiple dots with one dot.
- Tokenize the text in the two lists as follows: -
    - In the list of texts, if the last word ends with a dot (.). If a condition is true, the word is included in the list.
    - In the list of titles, append a dot (.) to the last word assuming that the last word is the end of a sentence title.
- Using the words module, which typically contains a list of correct English words from the NLTK library. Create a list (text-correct-word) to store the correct words after doing the following checks for each word:
    - It checks whether the last character of the word is a dot. If it is, it checks whether the stemmed version (using the stemming function) of the word without the dot itself, is in the word's module. If it is, the word is appended to the (text-correct-word) list.
    - Else, it checks whether the stemmed version or the word itself is in the words module and appends the word to the (text-correct-word) list if it is.
- Merge the two lists in one list.
- Applying lemmatization in all words in the new list.

- Produce a list of clear text.

B. Splitting and Proposed Reduction Method

The goal of this step is to reduce unimportant sentences in the document corpus by using the proposed (RS-LW) technique which includes the following steps:

- Using (NLTK. Tokenize. Sentence) for splitting the text into sentences based on the dot (.) and saving in the list of sentences.
- RS-LW reduces unimportant sentences in two ways:

o Computing the length of each sentence then removing the sentence which contains less than four words.

o Computing the weight for each sentence through:

- Compute the (TF-IDF) for each word in each sentence.
- Sum the (TF-IDF) for each words in each sentence.
- Add weight value (1.0) to each title sentence weight.
- Great new list (A) for the weights of each sentence divided by the number of words in each sentence.
sorting all sentences in descending order (selecting the first 75% from all sorted sentences).

C. Sentence Embedding using SBERT

Sentence embedding (SE) is the process of expressing a sentence in a continuous vector space as a fixed-size vector.
The objective is to effectively capture the sentence's semantic content for a variety of natural language processing (NLP) applications. A quick, small, and very efficient pre-trained SBERT model (the all-MiniLM-L6-v2 version) is used in this embedding procedure.
It has achieved state-of-the-art status by exhibiting exceptional performance in many phrase-embedding tasks [32].

D. Dimensionally Reduction using UMAP

Dimensionality reduction is a powerful tool in data analysis and machine learning that helps improve the efficiency and effectiveness of models, especially in scenarios were dealing with high-dimensional data is challenging. The UMAP is used to degrade the sentence embedding dimensionality that was obtained in section (Sentence Embedding).
Moreover, it assists in enhancing the performance of well-known clustering algorithms in terms of clustering precision and time. In this case, the parameters chosen for UMAP are (the number of neighbors and the number of components) in the lower-dimensional space, and the metric for computing distance (cosine similarity) [32, 33].

E. Topic Creation and Representation

Creating topic representations using C-TF-IDF (Class-based Term Frequency-Inverse Document Frequency) is an extension of the traditional TF-IDF algorithm. C-TF-IDF was applied to the sentences grouped by the HDBSCAN clustering algorithm to represent each topic with words as explained in the following steps and Table 1 [32]:

- Calculate TF for each term (word) in each sentence in each topic.
- Calculate IDF for each term in the other topics.
- Compute TF-IDF for each term by multiplying TF * IDF.
- Determine each topic with the top 10 words that represent the sentences included in it.

TABLE 1. (TOPICS NUMBER =12, WORDS NUMBER = 10 IN EACH TOPIC)

Topics NO.	Topic - Words
Topic 1	model, algorithm, data, function, learning, network, time, method, problem, matrix
Topic 2	bandit, armed, regret, problem, algorithm, setting, contextual, feedback, reward, bound
Topic 3	hashing, hash, hamming, distance, code, binary, function, loss, method, similarity
Topic 4	causal, graph, model, inference, discovery, effect, structure, causality, data, relationship
Topic 5	outlier, detection, anomaly, outliers, novelty, data, robust , method, point, algorithm
Topic 6	privacy, private, differentially, differential, algorithm, data, mechanism, user, output, bound
Topic 7	conference, international , proceeding, pages, machine, learning, mining, theory, annual, discovery
Topic 8	quantization, vector, error, data, tree, learning, product, performance, compression, method

Topic 9	copula, vine, model, bivariate, density, marginal, distribution, marginals, dependency, mixed
Topic 10	shot, zero, learning, class, training, classification, model, meta, task, unseen
Topic 11	spline, smoothing, knot, function, regression, basis, splines, kernel, cubic, model
Topic 12	odor, olfactory, bulb, cortex, receptor, neuron, activity, cell, pattern, input

5. THE PROPOSED TOPIC LABELING METHOD

The suggested method takes into account the lack of expression in the earlier methods and explains the proposed labeling method in the following steps:

- Bring the profiles from Google Scholar for all authors based on the name authors dataset.
  - utilizing the (Crossref Metadata API), which offers an API for scholarly content metadata access. Applications in Academia and research make extensive use of it.
- Extract the author's interests and save them in the list of interests after removing duplicate interests.
- Read the list of topics (Table 1).
- Apply S-BERT for encoding the two lists (interests, topics).
- Calculate the cosine similarity between topic embedding and all interests embedding, then assign the label for the topic by the max similarity value.
- Delete the interest from the list of interest and repeat the same step for all remaining topics.
- The Suitable Label for topics in Table 1 will be shown in Table 2.

TABLE 2. (TOPICS NUMBER =12, WORDS NUMBER = 10 IN EACH TOPIC, SUITABLE LABEL FOR EACH TOPIC)

Topics NO.	Topic Words	Suitable Label
Topic 1	model, algorithm, data, function, learning, network, time, method, problem, matrix	**Network Analysis**
Topic 2	bandit, armed, regret, problem, algorithm, setting, contextual, feedback, reward, bound	**Reinforcement Learning**
Topic 3	hashing, hash, hamming, distance, code, binary, function, loss, method, similarity	**Coding Theory**
Topic 4	causal, graph, model, inference, discovery, effect, structure, causality, data, relationship	**Graph Theory**
Topic 5	outlier, detection, anomaly, outliers, novelty, data, robust, method, point, algorithm	**Outlier Detection**
Topic 6	privacy, private, differentially, differential, algorithm, data, mechanism, user, output, bound	**Decentralized Optimization**
Topic 7	conference, international, proceeding, pages, machine, learning, mining, theory, annual, discovery	**Web Mining**
Topic 8	quantization, vector, error, data, tree, learning, product, performance, compression, method	**Machine Learning**
Topic 9	copula, vine, model, bivariate, density, marginal, distribution, marginals, dependency, mixed	**Probabilistic Modelling**
Topic 10	shot, zero, learning, class, training, classification, model, meta, task, unseen	**Transfer Learning**
Topic 11	spline, smoothing, knot, function, regression, basis, splines, kernel, cubic, model	**Differential Geometry**
Topic 12	odor, olfactory, bulb, cortex, receptor, neuron, activity, cell, pattern, input	**Sensory Systems**

6. EVALUATION METRICS

Topic coherence and variety indicators were used in conjunction with the recommended technique to assess the coherence and quality of the retrieved subjects.

The degree to which various words or even phrases within a topic fit inside the corpus is determined by topic coherence. Furthermore, it permits interpretability [19].

As CV (Co-Occurrence Value) seems to concur more with human assessment.

Based on the similarity between word pairings, this metric examines the relationships between each word, frequently with the use of vector space modules.
A value between 0 and 1 is typically used to express CV, the closer the value is to 1, the more consistent topic. The exact calculation of CV is formulated as [19,34]:

$$\mathrm{Cv}{=}v_{NPMI}(xi) = \{NPMI\ (xi, xj)\}_{\mathrm{j=1\ \ldots\ T}}$$

$$v_{NPMI}(\{xi\}_{i=1}^{T}\ ) = \{\textstyle\sum_{I=1}^{T} \mathrm{NPM}\ (xi, xj\}_{\mathrm{j=\ 1\ \ldots T}}$$

On the other hand, Topic Diversity represents a unique word ratio among the top terms from different topics. A value of diversity that is close to zero signifies redundant topics.

There measure of topic diversity: average Pairwise Jaccard. By multiplying the values of topic diversity and coherence, the overall quality of word groups occurring in each topic can be assessed [34].

7. RESULTS AND DISCUSSION

The "NIPS" dataset study shows that reduction sentences based on RS-LW are among the best performers when it comes to topic modelling. With its remarkable diversity, strong coherence, and excellent topic quality, RS-LW offers a sophisticated comprehension of the topic's matter.
Sentence embedding adds a high level of complexity by enabling the model to identify precise sentence details within topics. Researchers can use the proposed model (RS-LW) to find complex relationships and patterns, which in turn provides deep insights into the content of data and the dynamics of sentences.
The performance of (RS-LW) on the (NIPS) dataset is evaluated using evaluation metrics and then compared with other models, including LDA, ITMWE, ETM, and DTM. Table 3 shows the strong performance of the model (RS-LW) over another models.

The RS-LW model is better than the other models by gaining a high degree of Coherence score (0.59) as well as the differences in the topics obtained with a degree of Diversity score (0.96). In addition, the model obtained an excellent score in topic quality equal to (0.57), which was obtained from the product of multiplying the cohesion rate by the diversity rate. Table 4 shows Topic 1 obtained from NIPS dataset with top 10 words by the different models.

TABLE 3. PERFORMANCE OF THE MODEL (RS-LW) OVER ANOTHER MODELS WITH METRICS (TC, TD, TQ) [34].

Dataset	Method	TC	TD	TQ
NIPS	LDA (Latent Dirichlet Allocation)	0.3305	0.552	0.182
	ETM (Embedded Topic Model)	0.5340	0.732	0.390
	DTM (Dynamic Topic Model)	0.3011	0.183	0.055
	ITMWE (Incremental Topic Model with Word Embedding)	0.5766	0.884	0.509
	RS-LW (Reduction Sentences based on Length and Weight) (our proposed model)	**0.593**	**0.96**	**0.57**

TABLE 4. TOPIC 1 WITH TOP 10 WORDS BY THE DIFFERENT MODELS [35]

Method	LDA	DTM	ETM	ITMWE	RS-LW
Top-ten words for NIPS (Topic 1)	Character	Function	Structure	Model	**Model**
	Set	Class	covariance	Neural	**Algorithm**
	Training	Weight	Pattern	Function	**Data**
	High	Layer	Bias	Problem	**Function**
	Different	Use	Estimate	Set	**Learning**
	Network	Result	Noise	Result	**network**
	Learn	Time	Datum	Natural	**time**
	Sequence	Network	Condition	Datum	**method**
	Word	Training	Different	Network	**problem**
	Dimensionality	Example	Neural	Learn	**matrix**

8. CONCLUSION

Scientific publications emphasize the importance of analysis, which in turn helps understand complex textual. This study sheds light on this issue by using the (RS-LW) model, which relies on reducing unimportant sentences based on the length and weight of the sentence, also give more weights of each title for all papers, thus improving the topic modeling process.

Sentence-BERT used for sentence embedding process, as well as UMAP used for dimensionality reduction, and finally utilize the HDBSCN for clustering are all used with (RS-LW) respectively in order to obtain more coherence topics.

From the experiments in above table 3 conclude that The scores of (0.593) and (0.96) for topic coherence and diversity, respectively are the best with (RS-LW) approach.

Also, the topics generated from topic modelling process lacks to labels which are important for researcher to understands the textual more clearly, to tackle this issue, the study presents the unique technique of keyword generation using the extraction of interests of all authors from authors' Google Scholar profiles.

By automating the tagging process, this method seeks to lessen the cognitive load and enable a more insightful analysis of the subjects that are extracted from the text corpus.

Through the experiments show that the obtained keywords labels are very suitable and matching with the resulted topics. All things considered, the study advances topic modelling methods and offers novel approaches to subject interpretation and labelling in the context of scientific publications.

9. REFERENCES

[1] Dieng, Adji B., Francisco JR Ruiz, and David M. Blei. (2020) "Topic modeling in embedding spaces." Transactions of the Association for Computational Linguistics 8, 439-453.

[2] Al-Tai, Mohammed Haqi, Bashar M. Nema, and Ali Al-Sherbaz. (2023) "Deep learning for fake news detection: Literature review." Al-Mustansiriyah Journal of Science 34.2, 70-81.

[3] Shaker, N.H., Dhannoon, B.N. (2024) "Word embedding for detecting cyberbullying based on recurrent neural networks" 13(1), 500–508.

[4] Salman, Zainab Abdul-Wahid (2023). "Text Summarizing and Clustering Using Data Mining Technique." Al-Mustansiriyah Journal of Science 34.1 58-64.

[5] QIANG, Jipeng, Et Al. (2017) Topic Modeling Over Short Texts by Incorporating Word Embeddings. In: Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21. Springer International Publishing, p. 363-374.

[6] Wotaifi, T.A., Dhannoon, B.N. (2023) "Developed Models Based on Transfer Learning for Improving Fake News Predictions." JUCS: Journal of Universal Computer Science 29.5 491-507.

[7] Mann, Jasleen Kaur. (2021) "Semantic Topic Modeling and Trend Analysis.".

[8] Jelodar, Hamed, et al. (2019) "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." Multimedia Tools and Applications 78, 15169-15211.

[9] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003) "Latent Dirichlet Allocation." Journal of machine Learning research 3. Jan, 993-1022.

[10] Sbalchiero, Stefano, and Maciej Eder. (2020) "Topic modeling, long texts and the best number of topics. Some Problems and Solutions." Quality & Quantity 54, 1095-1108.

[11] Avasthi, Sandhya, Ritu Chauhan, and Debi Prasanna Acharjya. (2021) "Processing large text corpus using N-gram language modeling and smoothing." Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI 2020. Springer Singapore.

[12] Ibrahim, Mohammed F., Mahdi Ahmed Alhakeem, and Nawar A. Fadhil. (2021) "Evaluation of Naïve Bayes classification in Arabic short text classification." Al-Mustansiriyah J. Sci. 32.4, 42-50.

[13] Peters, Seth. (2022) "Bringing Visibility to Workflow Processes via Topic Modeling with BERT Transformer Models". MS thesis. Itä-Suomen yliopisto.

[14] Zhou, Mei; Kong, Ying; Lin, Jianwu. (2022) "Financial Topic Modeling Based on the BERT-LDA Embedding". In: 2022 IEEE 20th International Conference on Industrial Informatics (INDIN). IEEE, p. 495-500.

[15] HRISTOVA, Gloria; NETOV, Nikolay. (2022) Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic:" A Topic Modeling Approach Based on BERTopic". In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 2259-2264.

[16] Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022) " Topic discovery via latent space clustering of pretrained language model representations". In Proceedings of the ACM Web Conference, pp. 3143-315.

[17] Oyshi, Uttamasha Anjally. (2023) "Topic Modeling and Prediction of Aid Data in Development Studies Using LDA and BERT". Diss. University of Arkansas at Little Rock.

[18] Kozbagarov, Olzhas; Mussabayev, Rustam; MLADENOVIC, Nenad. (2021) " A new sentence-based interpretative topic modeling and automatic topic labeling". Symmetry, 13.5: 837.

[19] Avasthi, Sandhya; Chauhan, Ritu; Acharjya, Debi Prasanna. (2021) "Processing large text corpus using N-gram language modeling and smoothing". In: Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI 2020. Springer Singapore, pp. 21-32.

[20] Scarpino, Ileana, Et Al. (2022) "Investigating Topic Modeling Techniques to Extract Meaningful Insights" in Italian Long Covid Narration. Biotech, 11.3: 41.

[21] Yang, Xi, Et Al. (2022) "A Study of Text Vectorization Method Combining Topic Model and Transfer Learning". Processes, 10.2: 350.

[22] George, Lijimol, and P. Sumathy. (2023) "An integrated clustering and BERT framework for improved topic modeling." International Journal of Information Technology 1-9.

[23] Ogunleye, Bayode, Et Al. (2023) "Comparison of Topic Modelling Approaches in The Banking Context". Applied Sciences, 13.2: 797.

[24] Talebpour, Mozhgan; García Seco De Herrera, Alba; Jameel, Shoaib. (2023) " Topics in Contextualised Attention Embeddings". In: European Conference on Information Retrieval. Cham: Springer Nature Switzerland, P. 221-238.

[25] Tang, Guofeng, Et Al. (2023) "Research on The Evolution of Journal Topic Mining Based on The Bert-Lda Model". In: Shs Web of Conferences. Edp Sciences, P. 03012.

[26] Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. *Natural Language Processing Journal*, *6*, 100053.

[27] Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. (2016) "Automatic labelling of topics with neural embeddings." arXiv preprint arXiv:1612.05340.

[28] Gourru, Antoine, Et Al. (2018) United We Stand: "Using Multiple Strategies for Topic Labeling. In: Natural Language Processing and Information Systems": 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23. Springer International Publishing, P. 352-363.

[29] He, Dongbin, Et Al. (2021) "Automatic Topic Labeling Using Graph-Based Pre-Trained Neural Embedding". Neurocomputing, 463: 596-608.

[30] Avasthi, Sandhya, and Ritu Chauhan. (2024) "Automatic label curation from large-scale text corpus." Engineering Research Express.

[31] Tang, X., Mou, H., Liu, J., & Du, X. (2021). Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports*, *11*(1), 11849.

[32] Sawant, Sahil, Et Al. (2022)."An Enhanced Bertopic Framework and Algorithm for Improving Topic Coherence and Diversity". In: 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems&Application(HPCC/DSS/Smartcity/Dependsys). IEEE, P. 2251-2257.

[33] Gelar, Trisna; SARI, Aprianti Nanda. (2024) "Bertopic and NER Stop Words for Topic Modeling on Agricultural Instructional Sentences". In: International Conference on Applied Science and Technology on Engineering Science 2023 (iCAST-ES 2023). Atlantis Press, p. 129-140.

[34] Grootendorst, Maarten. (2022) "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." arXiv preprint arXiv:2203.05794.

[35] Avasthi, Sandhya, Ritu Chauhan, and Debi Prasanna Acharjya. (2023) "Extracting information and inferences from a large text corpus." International Journal of Information Technology 15.1 435-445.

Rana F. Najeeb

I am PHD Student in the last step of the research. My main researches concerns are: 1. Artificial Intelligent (machine learning, Deep Learning) 2. Digital Image Processing, 3. Data Mining ,4. Classification 5. Clustering.

Ban N. Dhannoon

I am Ban N. Dhannoon, PhD holder since 2001, currently, a member of teaching staff in Computer Science Dept. / College of Science/ Al-Nahrain University, Iraq. My main researches concerns are: 1. Artificial Intelligent (machine learning, Multi-agent, Deep Learning) 2. Digital Image Processing 3. Coding (encryption, data compression, representation) 4. Pattern Recognition & Classification 5. Bioinformatics.

Farah Qais Alkhalidi

Assistant Prof. in computer science, Mustansiriyah University. My main researches are: 1. Processing 2. Information Technology 3. Graphics 4. thermal imaging.