IRAQI Academic Scientific Journals

Samarra Journal of Pure and Applied Science

SJPAS

مجلة سامراء للعلوم الصرفة والتطبيقية

www.sjpas.com

p ISSN: 2663-7405
e ISSN: 2789-6838

From Data to Insight: Topic Modelling and Automatic Topic Labelling Strategies

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

1- Department of Computer Science, College of Science, Mustansiriyah University, Iraq
2- Department of Computer Science, College of Science, Al-Nahrain University, Iraq

https://doi.org/10.54153/sjpas.2025.v7i4.1061

Article Information

Received: 21/10/2025
Revised: 24/11/2025
Accepted: 28/11/2025
Published: 30/12/2025

Keywords:

Deep Learning, S-BERT, Dimensionally Reduction, Topic Coherent, and Topic Diversity.

Corresponding Author

E-mail:
rana19.najeeb@uomustansiriyah.edu.iq

Mobile: 07700013960

Abstract

In order to enhance the interpretability of data for decision-making, scientific, biological, and social media text collections require efficient machine learning techniques. Text mining is aided by topic models in sources such as blogs, Twitter data, scientific journals, and biomedical papers. It is still difficult to find appropriate labels, even when topic modeling indicates important concepts. Analysts' cognitive effort is decreased by automating topic evaluation and categorization. While certain techniques rely on word frequency to produce labels with words, phrases, or images, extractive methods choose labels based on probability measures. This study suggests improving the topic modeling in a collection of conference papers on Neural Information Processing Systems (NIPS) released between 1987 and 2017 and achieved two goals: producing more coherent topics and automatic topic labelling. The first goal was achieved through five phases: text pre-processing phase, reduction phase using a new method called SR-LW (Sentences Reduction Based on Length and Weight), which removes sentences of shorter length, then calculates the weight for the remaining sentences and removes approximately 25% of the less weight sentences. The sentence embedding phase uses S-BERT (Sentence-Bidirectional Encoder Representation from Transformer) to reduce the dimensionality of the sentence embedding phase by utilizing the Uniform Manifold Approximation and Projection (UMAP). Lastly, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) organized comparable documents. The experimental findings demonstrate that using the proposed SR-LW phase has produced more cohesive topics, improving topic coherence by (0.593) and topic diversity performance by (0.96). Though topic modelling 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' profiles in Google Scholar and extracting the interests for use in automatically labelling the topics.

Introduction

Communication technology has changed and transmitted information rapidly over the past years, as manifested by the rise of social media. News headlines, tweets, social media posts, blog entries, user comments, news stories, scientific articles, and many more are just a few sources that provide textual information [1, 2]. The information can be, by definition, social media, scientific, biological, or operational, demonstrating the diversity of various datasets [3]. Among others, efficient machine learning methods are needed to extract inherent themes or topics in the accurate interpretability of data and clustering information to meet the objectives [4]. The popular approach of this kind is the topic model; topic modeling is an approach to unsupervised machine learning since topics and mixture parameters are not known but are purely inferred from data [5, 6]. In other words, it is not trained on already tagged or labeled data [7]. The most commonly used probabilistic topic modeling technique is Latent Dirichlet Allocation (LDA) [8], and another very foundational topic modeling technique is probabilistic latent semantic analysis (pLSA) [8],[9] 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 called the Bidirectional Encoder Representation from Transformer (BERT) model based on neural networks [10]. The benefit of using BERT is that the model assigns a vector to the whole sequence or sentence in the dataset [11], taking the averaged embedding of its constituent words [12]. An essential task for a topic model is labeling topics. This is an algorithmic process for generating/selecting the best-formed phrases or sentences that describe the topic. This field of research needs to be better researched and developed [13]. One of the earliest techniques for labelling text was hand labelling topics. The manual labeling of themes was a key component of specific previous labeling strategies. In the case of manual tagging, a competent user assigns a tag that captures the topics by observing a set of words concerning a given problem. This manual method is time consuming and intensive in human interactions [14].

The objectives of topic modeling and labeling on the NIPS dataset include; increased comprehension and organization of a large collection of research papers in the fields of machine learning and neural information processing, improved information search, collaboration, uses in machine learning and decision making as well as advancing the fields academically and industrially. In line with topic modeling, the objective of the study is to obtain the abstract and meaningful representation of the given dataset to analyze the dynamics of the research trend. This understanding helps in the determination of the future research agenda and reduces research duplication, thus improving the efficiency and effectiveness of the research. The benefits of topic labeling include the ability to properly organize the dataset and make it easy for the researchers to locate the papers of interest. It also enhances the effectiveness of information retrieval systems primarily in order to increase the relevancy and relevance of the returned results to the user’s query. This can aid researchers in locating the most pertinent studies and resources quickly, enhancing their work.

Contribution of the Study is as follows:

- The study propose a brand-new technique called Sentence Reduction Based on Length and Weight (SR-LW), which removes shorter, less significant sentences in order to improve topic modeling. The quality of topic extraction is enhanced and noise is decreased with this method. More cogent and varied topics are covered in the study by combining SR-LW with sophisticated tools like S-BERT (for sentence embeddings), UMAP (for dimensionality reduction), and HDBSCAN (for clustering).
- Propose a new system that automatically generates labels to address the problem of topic labeling Using interests extracted from authors' Google Scholar profiles as keywords for the topics,. The labeling procedure is more accurate and efficient using this method since it drastically lowers the cognitive load on analysts.

This study is organized into five sections. The first section introduces the topic of modeling and labeling problem statements and their importance. The second section covers research on different topics, such as modeling and labeling algorithms. In the third section, the description of the dataset used is elaborated. Proposed topic modeling and labeling algorithms are discussed in sections 4 and 5, respectively. Further, Section 6 gives the evaluation, Section 7 presents the results obtained, and the conclusion is given in Section 8.

Related Work

The Topic Modelling 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 [15]. Topic labeling is to generate semantically appropriate text categorization or word group labels automatically. This literature review provides an in-depth examination of studies dealing with extracting labels and topics from corpus or text collections [16, 17].

1. Topic Modelling

The specifics of the proposed approach involve assessing sentence probabilities within the text corpus and clustering sentence embedding. The thesis in [5] delineates a method for generating semantically meaningful topics by leveraging contextual embeddings like BERT and Sentence-BERT. In [18], topic clustering is developed based on the BERTLDA joint embedding, considering contextual semantics and the topic narrative. Cluster text embedding through the HDBSCAN algorithm and c-TF-IDF technique is used to build the topic representation. The BERT-LDA model is reliant on high-quality data, needs a lot of resources, and has trouble with interpretability and short sentences. Its performance differs depending on the domain and requires improvement for wider use. The study in reference [19] centres on applying Transformer models, such as 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 [20] combines transfer learning with a topic model. First, to model the data from the text and extract its keywords, the topic model is chosen to identify the primary information included in the data. Next, model transfer learning is performed to produce vectors to compute text similarity between texts using the pre-trained models (The BERT algorithm) model. In [21], a study that combines the benefits of BERT and LDA topic modeling techniques presents a unified clustering-based framework for mining significant themes from massive text corpora. The study draws attention to several drawbacks, including restricted dataset generalizability, computational cost, and dependence on dimensionality reduction. It recommends using more sophisticated models in the future to enhance subject modeling performance. The paper in [22] describes how the BERTopic architecture uses K-means clustering and Kernel Principal Component Analysis (Kernel PCA).

The primary goal of [23] was to identify word topic clusters in pre-trained language models using the BERT and Distil-BERT. The attention framework was discovered to be crucial to modeling this kind of word topic clustering. In [24], the authors proposed an LDA-Bert public opinion topic mining model to solve the problem that LDA ignores context semantics and the topic distribution is biased towards high-frequency words. The manual labeling process is labor-intensive, posing challenges for scalability with larger datasets. The most important of this study [25] is to identify the types of information and their effective impact on an analytical theory for analyzing fake news related to the Coronavirus (Covid-19). It merges the idea of sentiment analysis with the Topic Model for source optimization of copious amounts of unstructured material through viewing the feelings of words. The dataset contains 10,254 custom addresses from all over the world. The study has several drawbacks, such as the work-intensive process of manually classifying topics, possible biases from certain data sources, and context-dependent performance that could restrict generalizability. It demands a lot of computing power and concentrates on coherence scores, which might not accurately represent actual utility, making it difficult for wider application.. Among the models tested, LDA showed the best acceptable rate: 0.66 for 20 items that resulted in affected feelings and 0.573 for 18 false positives, resulting in Subjects that excel in NMF at the p-value of 0.43 and in LSA at the p-value of 0.40.

2. Topic Labelling

Topic Labelling Under some guidance, [26] recommends providing a topic with a succinct label that captures its principal concept or topic. Using Wikipedia article titles as label possibilities, the neural embedding for both words and documents was computed to choose the finest labels for the subjects. The authors of [27] presented the fourth topic labeler, the representative sentence extractor with Dirichlet smoothing. This sentence-based labeler provided good surrogate candidates in cases where the n-gram topic labelers fell short in giving relevant labels, with up to 94The study in [28] presented a unique two-phase neural embedding system incorporating a graph-based ranking method mindful of redundancy. It illustrated how applying pre-trained neural embedding could benefit topic names, sentence demonstrations, and tasks of automatic topic labeling.

The approach described in [14] allows for automatically generating labels to represent every topic. It uses a labeling strategy in which the candidate labels are filtered, and then sequence-to-sequence labelers are applied. This approach aims to obtain a meaningful label for the output of LDA algorithms. The limitations of topic labeling techniques include domain-specificity, manual intervention, generic labels, lack of context, biases in AI systems, and overlapping topics. These difficulties show that more precise and flexible methods are required. The article in [29] 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 several suitable stickers. Finally, the word2vec model was used for Deep text analysis.

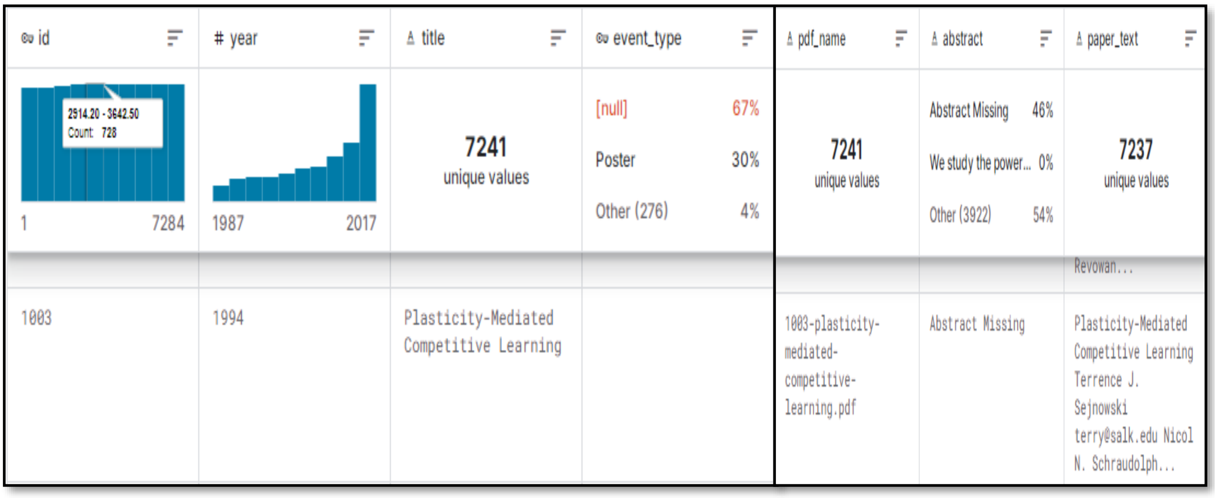
Dataset Description:

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 [1], [14]. 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, full text and units. NIPS consist of four files (authors.csv, database.sqlite, paper_authors.csv, and papers.csv). In this study the files used are two files as shown in Figure 1 ((a) authors.csv and (b) papers.csv).

1. NIPS - Papers.csv

   This is the primary file that contains metadata and full text about the NIPS papers.

   - ID: A unique identifier for each paper.
   - Title: The title of the paper.
   - Abstract: A brief summary or abstract of the paper's content.
   - Year: The year of the conference when the paper was presented.
   - PDF name: The filename or link to the full paper in PDF format.
   - Event type: Type of event the paper was presented at, such as a workshop or poster session.
   - Full Text: Sometimes included as a separate column, this contains the full text of the paper.



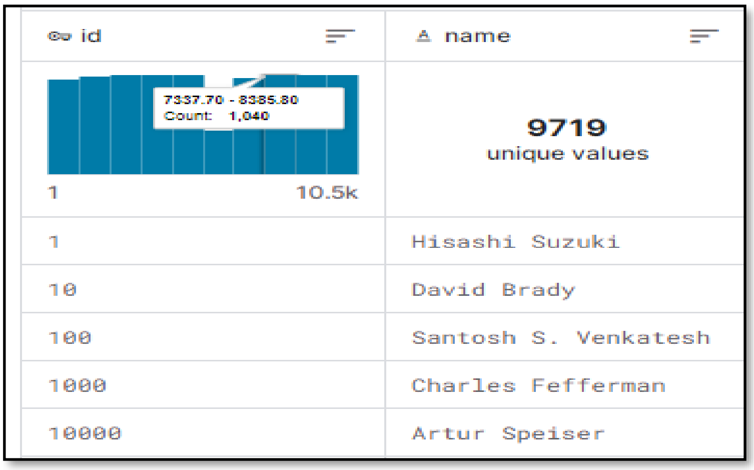
id	# year	title	event_type	pdf_name	abstract	paper_text
2914.20 - 3642.50 Count: 728 1 7284	1987 2017	7241 unique values	[null] 67% Poster 30% Other (276) 4%	7241 unique values	Abstract Missing 46% We study the power... 0% Other (3922) 54%	7237 unique values
						Revowan...
1003	1994	Plasticity-Mediated Competitive Learning		1003-plasticity-mediated-competitive-learning.pdf	Abstract Missing	Plasticity-Mediated Competitive Learning Terrence J. Sejnowski terry@salk.edu Nicol N. Schraudolph...

Fig. 1 (a) NIPS-Ppapers.csv.

2. nips_authors.csv

   The file contains the information about the authors of the NIPS papers.

   - Author ID: A unique identifier for each author.
   - Author name: The name of the author.



id	name
7337.70 - 8385.80 Count: 1,040 1 10.5k	9719 unique values
1	Hisashi Suzuki
10	David Brady
100	Santosh S. Venkatesh
1000	Charles Fefferman
10000	Artur Speiser

(b): NIPS- Authors File.csv

Methodology

This study create a reliable workflow for in-depth topic modeling by utilizing cutting-edge methods like Sentence BERT, UMAP, and HDBSCAN to analyze conference papers from Neural Information Processing Systems (NIPS), as shown in Figure 2. It stresses the importance of the recently released SR-LW algorithm, which is intended to analyze big text datasets effectively, and contrasts it with more conventional approaches. Key shortcomings of current methods are addressed by the SR-LW algorithm, which stands out for its capacity to improve topic modeling by lowering noise and enhancing data coherence.

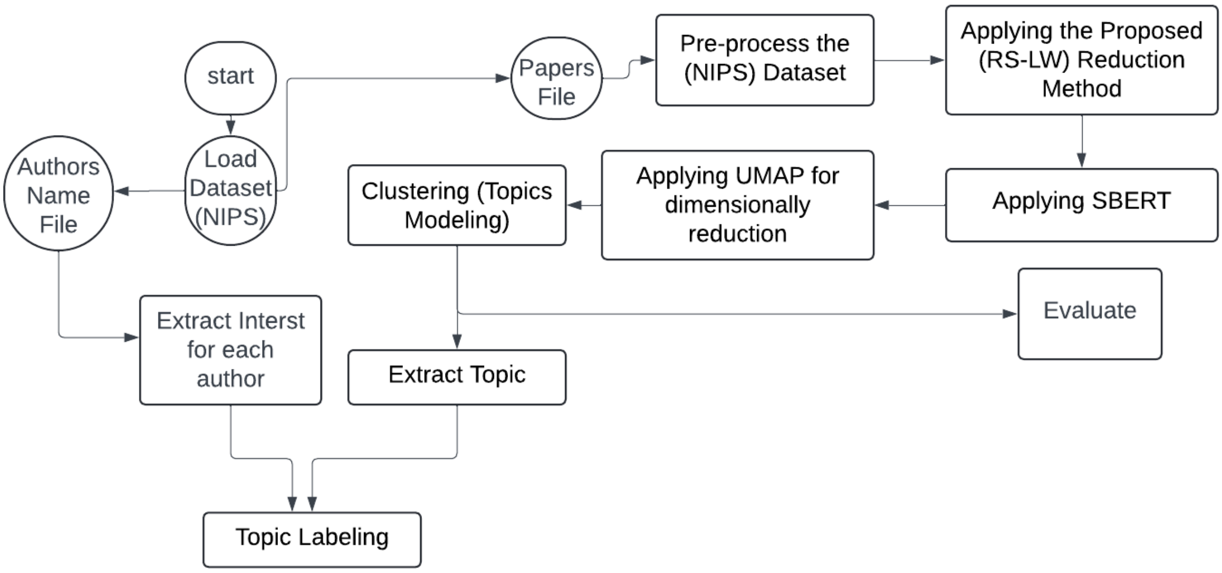

start
Load Dataset (NIPS)
Authors Name File
Papers File
Pre-process the (NIPS) Dataset
Applying the Proposed (RS-LW) Reduction Method
Applying SBERT
Applying UMAP for dimensionally reduction
Clustering (Topics Modeling)
Evaluate
Extract Interst for each author
Extract Topic
Topic Labeling

Fig. 2 Flowchart of Work Steps

1. Data Pre-processing

Data cleaning is necessary since text data can have different formats and contain errors. Before begin the pre-processing steps, there are some challenges in NIPS dataset must be handle with it as explained in the listed points:

- Create list of words is equivalent in importance to standard stop words but are especially for NIPS data set in order to removed it.

- Using the words module, this typically contains a list of correct English words from the NLTK library, to check the corrected words.
- Deals with the period in different steps were it necessary when break the text into sentences.

The following steps are taken to achieve a targeted text format for each paper in the dataset as shown in Figure 3:

- Creating two blank lists (List-of-papers and List-of-Titles).
- For the text papers and title, do the following:
  - Step 1: Convert to lowercase.
  - Step 2: Remove Digits and Special Characters. Place before dots at the end of sentences, stop words, and custom special words and words less than three characters.
- Additional steps for Text paper
  - Step 3: Remove the text between the title and the abstract, the new acknowledgments, and references.
  - Step 4: Removing incorrect words (Wi: word) (by checking if the stemming of (Wi) was not found in the list of correct words.
- Added the title to the list of title.
- Added the title to the text paper.
- Applying Lemmatization.
- Added paper text to the list of text.

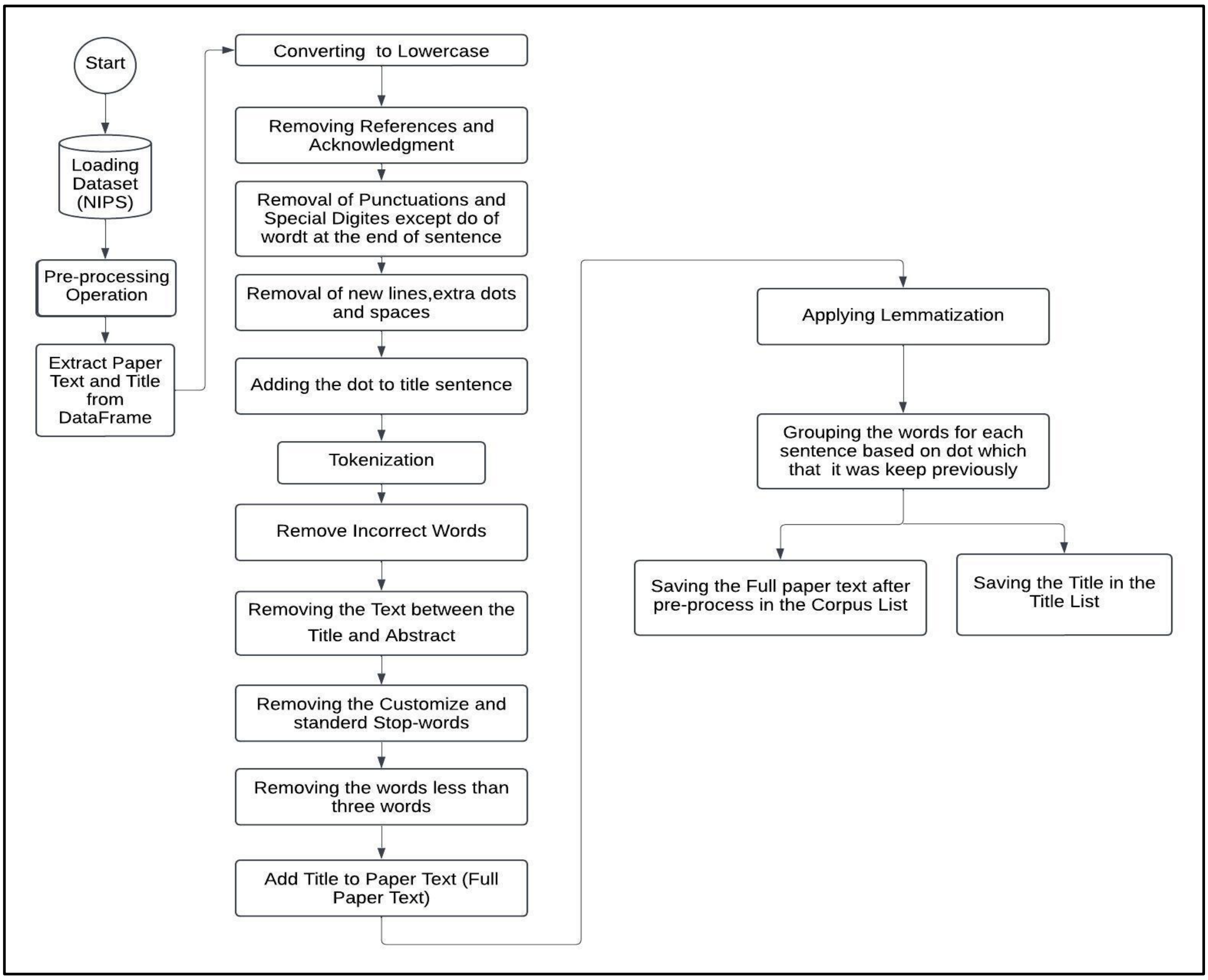

Start
Loading Dataset (NIPS)
Pre-processing Operation
Extract Paper Text and Title from DataFrame
Converting to Lowercase
Removing References and Acknowledgment
Removal of Punctuations and Special Digites except do of wordt at the end of sentence
Removal of new lines,extra dots and spaces
Adding the dot to title sentence
Tokenization
Remove Incorrect Words
Removing the Text between the Title and Abstract
Removing the Customize and standerd Stop-words
Removing the words less than three words
Add Title to Paper Text (Full Paper Text)
Applying Lemmatization
Grouping the words for each sentence based on dot which that it was keep previously
Saving the Full paper text after pre-process in the Corpus List
Saving the Title in the Title List

Fig.3 Pre-processing Steps

2. Feature Extraction

It plays a crucial role in preparing the data and enhancing the effectiveness of clustering algorithms. The goal is to select or create features that highlight the underlying patterns in the data, making it easier for clustering algorithms to group similar data points together. The utilized textual feature extraction approaches are discussed below [30]: -

a) TF-IDF (Term Frequency-Inverse Document Frequency)

- TF is a statistical quantity that assesses how pertinent a word is to a document in a group of documents. So, two things are mentioned. These are the inverse document frequencies and the term [31].
- IDF (Inverse Document Frequency) is used to calculate the weight of rare words across all documents.
- TF-IDF is computed by multiplying the frequency of the term in the document by its inverse across a collection of documents.
- C-TF-IDF (Class-Based TF-IDF) A variation of the traditional Term Frequency-Inverse Document Frequency (TF-IDF) technique that considers each document's class or category information. It is commonly used in machine learning and natural language processing tasks where documents belong to distinct classes or categories. The objective is to calculate the importance of terms not just within individual documents but within the context of specific classes or groups [32].

b) Transformer-based Models

All the NLP models are language models based on transformers. The transformer consists of two main components: an encoder and a decoder. The encoder takes words as input to generate embedding that encapsulates the word's meaning. In contrast, the decoder takes the embedding generated by the encoder to generate the next word until the end of the sentence. One of the transformer-based models is named BERT and S-BERT [33].

- BETR (Bidirectional Encoder Representations from Transformers): This is a sentence encoder to derive a contextual representation of sentences correctly given. BERT overcomes the unidirectionality constraint through masked language modeling. During masked language modeling, it masks several tokens from input at random, using the input to guess the original vocabulary ID of the masked word [32].
- S-BERT (Sentence-Bidirectional Encoder Representations from Transformers) is a modified model of BERT designed to generate high-quality sentence embeddings. It encodes entire sentences or paragraphs into dense vectors, enabling efficient semantic search by comparing these vectors for similarity. This capability makes S-BERT ideal for tasks like clustering similar documents and topic modeling due to their ability to capture semantic meaning effectively. It's also been applied in other works like semantic textual similarity, semantic search, and paraphrase mining via a Siamese network with a pre-trained BERT model. The Siamese network architecture enables fixed-sized vectors to be derived from input sentences. Pooling is a mechanism to obtain a fixed-size vector representation for the entire

input sequence, whether a sentence, paragraph, or document. Pooling is applied after the transformer model has generated the token embeddings. Semantically similar sentences can be found using a similarity measure like cosine-similarity or Manhatten / Euclidean distance [32].

3. The Proposed Reduction Algorithm

The Sentence Reduction Based on Length and Weight (SR-LW) technique is essential for improving topic modeling since it eliminates unnecessary sentences from a corpus of documents. The dataset is optimized for quicker and more efficient processing while improving topic coherence and diversity by eliminating noise and keeping only the most pertinent information. This approach offers higher scalability and reduces unnecessary information, making it especially useful for huge datasets. In the topic modeling process, SR-LW is a crucial and effective phase that improves the performance of sophisticated methods like Sentence-BERT, UMAP, and HDBSCAN by offering an improved dataset. The SR-LW includes the following steps as shown in Algorithm 1:

```
Algorithm 1 : Proposed Reduction Algorithm (SR-LW)

Input: Full Papers Text (after pre-processing).
Output: Reduced Dataset (Papers Text).

Preparing step: Break Down Papers Text into sentences (S).
For each S:  i =1 to N : Number of Sentences do the following :
   For j= 1 to M : Number of Words (W) in S do the following:
Step1: Compute the (TF-IDF) which explained in section (2.6) in chapter two for each Wj in each Si.
Step2: Sum the computed (TF-IDF) for each Wj in each Si.
Step3: Before the dividing process checking:
if Si is a title sentence in the second list (Title List) which is prepare previously then
    -  Add weight value about (1.0) to the (TF-IDF) summation of the title sentence words because it a pivotal in capturing attention, clarifying the subject matter.
Step 4: Compute Averaging Weight (AW) For each S by Equation (1):
                              $AV= Sum / M$…… (1)
End For
  End For
Step 5: Sorting all sentences according to its averaging weight from high to low and then delete the fewer weights value about 25% of sentences to keep the most sentences (70%) which are considered the important in the analysis.
Step 6: End.
```

4. Sentence Embedding using S-BERT

Sentence embedding (SE) expresses a sentence in a continuous vector space as a fixed-size vector. This step aims to effectively capture the sentence's semantic content for various natural language processing (NLP) applications. A quick, small, and efficient pre-trained Sentence Bidirectional Encoder Representation from Transformer (SBERT) model (the all-MiniLM-L6-v2 version) is used. The"all-MiniLML6-v2" model is a variant of the MiniLM model, a compact and efficient language model designed for various natural language processing tasks. It's characterized by its small size and fast inference speed, making it suitable for applications where computational resources are limited or where quick responses are necessary [34].

5. Dimensionally Reduction using UMAP

Reducing the dimensions is one of **the** most effective techniques in data analysis and machine learning that can enhance the performance of models in situations when it becomes problematic to work with high-dimensional data. It is particularly valued for its ability to preserve the global structure and local relationships of data better than many other techniques. The UMAP is used to degrade the sentence embedding dimensionality obtained in the section after embedding the sentences. It assists in enhancing the performance of well-known clustering algorithms in words of clustering precision and time. Its efficiency and scalability make it suitable for large datasets, offering meaningful insights and patterns that are otherwise difficult to discern in high-dimensional space. In this case, the parameters chosen for UMAP are (the number of neighbours, the number of components) in the lower-dimensional space, and the metric for computing distance (cosine similarity) [35].

6. Topic Creation and Representation

Topics were generated using the Hierarchical Spatial Clustering for Noisy Applications (HDBSCAN) algorithm, where each obtained topic contains a set of sentences. HDBSCAN operationally determined the number of subjects based on the dataset used. The number of topics here is (12) topics. Category-based inverse document frequency (CTF-IDF), an extension of the traditional TF-IDF algorithm, was applied to the sentences collected by the HDBSCAN algorithm to represent each topic with the top 10 words. The steps for C-TF-IDF calculations are explained in the following points and as shown in Table 1:

- Calculate the TF for each word in each sentence in each topic.
- Calculate IDF for the same word in the other topics.
- Compute TF-IDF for each word by multiplying TF by IDF.
- Determine each topic with the top 10 words representing the sentences included.

Table 1: Top 10 Words in Each Topic

Topic 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

7. Evaluation Metrics

In topic modeling, coherence and diversity are crucial metrics that guarantee subjects that are both meaningful and interpretable while capturing a wide variety of ideas. Whereas diversity guarantees unique and thorough coverage, coherence concentrates on semantic similarities within themes. When combined, they offer a fair assessment of topic quality, which makes them perfect for practical uses. Topic coherence and variety indicators were used with the recommended algorithm to assess the coherence and quality of the retrieved topics. Topic coherence determines the degree to which various words or phrases within a topic fit within the corpus. Furthermore, it permits interpretability [1]. CV (Co-Occurrence Value) concurs more with human assessment. Based on the similarity between word pairings, this metric frequently examines the relationships between each word using 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 the topic. The exact calculation of CV is formulated as [1]:

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

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

On the other hand, Topic Diversity represents a unique word ratio among the top words from different topics. A value of diversity that is close to zero signifies redundant topics. There is a measure of topic diversity: the proportion of unique words (puw), which measures the variety of vocabulary in a given text. It is defined as the ratio of unique words to the total number of words in the text. By multiplying the values of topic diversity and coherence, the overall quality of word groups occurring in each topic can be assessed [1].

8. Results and Discussion

Reduction sentences based on SR-LW are among the best at topic modeling, according to the "NIPS" dataset analysis. SR-LW offers a deep comprehension of the subject matter and exhibits remarkable diversity, strong coherence, and good topic quality. The proposed SR-LW model is the best among the rest, as it has earned a highly coherent score of (0.59) and the differences in topics, with a high degree of diversity, of (0.96). Moreover, the model scored well regarding topic quality, equal to (0.57), which was earned by multiplying the cohesion rate by the diversity rate. Table 2 shows how much the improvement in the model's performance which using HDBSCAN clustering algorithm by applying the proposed reduction Technique (SR-LW).

Table 2: The Performance Comparative of SR-LW

Technique	NO. of Sen.	TC-CV	TD-PUW	TQ	Running Time (Sec.)
Without SR-LW+ HDBSCAN	1391063	0.5	0.682	0.341	1082.9
With SR-LW + HDBSCAN	685500	**0.593**	**0.944**	**0.559**	**136.66**

This shows the improvement in the results by applying the proposed technique (SR-LW). It also shows the time when implementing the HDBSCAN clustering algorithm with and without using the proposed (SR-LW). Table 3 shows Topic 1 obtained from the NIPS dataset with the top 10 words by the different models in [1] and the proposed SRLW model.

Table 3: Top 10 Words in NIPS for Topic 1 Suggested by Different Models

Method	LDA	DTM	ETM	ITMWE	SR-LW
Top-ten words for NIPS (Topic No. 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**

The performance of (SR-LW) on the (NIPS) dataset is evaluated and then compared with other models, including LDA, ITMWE, ETM, and DTM. Table 4 shows the comparison the proposed SR-LW model with HDBSCAN with other different model in the NIPS dataset.

Table 4: Comparison our model with Different Models [1]

REF.	Method	TC	TD	TQ
[1] Whole NIPS	LDA (Latent Dirichlet Allocation)	0.331	0.552	0.182
	ETM (Embedded Topic Model)	0.534	0.732	0.390
	DTM (Dynamic Topic Model)	0.301	0.183	0.055
	ITMWE (Incremental Topic Model with Word Embedding)	0.577	0.884	0.509
Our model	**SR-LW + HDBSCAN**	**0.593**	**0.96**	**0.57**

9. Automatic Topic Labeling:

After the topic modeling or clustering process, the top 10 terms representing each cluster's topic are identified. However, these term lists are often insufficient for human interpretation, highlighting the need for effective topic labeling. To address this, we propose a new algorithm for topic labeling designed to enhance the accuracy and interpretability of automatically generated labels as shown in Table 5. Unlike traditional keyword-based methods that often produce generic or ambiguous labels; our approach considers the broader context of the topic's top words, ensuring the labels capture the true meaning and nuances of the topics. The steps of the proposed algorithm are detailed below and illustrated in Algorithm 2:

```
Algorithm 2 : Proposed  Automatic Topic Labeling Algorithm
Input: Generated Topics + Author Name File in NIPS Dataset.
Output: Suitable Label for each topic.

Step 1: Using author name with removing the redundant name to bring the profile
        information for each author from the Google scholar using Google Scholar API.
Step2: Extract the interest for each author from its profile.
Step 3: Embedding (Topics and Interests) by applying SBERT algorithm.
Step 4: For each Topic Compute:
    ➢ The similarity with all interests.
    ➢ Assign the higher value similarity for that interest to be considering the suitable
      label for this topic.
    ➢ Remove this interest and repeat those processes for all remaining topics and
      interests.
Step 5: End.
```

Table 5: Automatic Label for Extracted Topics

Topic NO.	Topic Words	Topic NO.
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**

10. Conclusion

Scientific publications highlight the critical role of analysis in understanding complex text. This study addresses this challenge through the SR-LW (Sentence Reduction Based on Length and Weight) model, which prioritizes important content by filtering out less relevant sentences based on length and weight. The SR-LW model also assigns higher weights to paper titles, enhancing the overall topic modeling process. Integrated with advanced techniques like Sentence-BERT for sentence embedding, UMAP for dimensionality reduction, and HDBSCAN for clustering, the SR-LW algorithm generates topics with greater coherence and diversity, achieving scores of 0.593 and 0.96, respectively, as shown in Table 2.

Despite these advancements, the study identifies the lack of topic labels as a barrier to fully understanding the extracted themes. To address this, it introduces a novel algorithm for automatic keyword generation, leveraging the research interests extracted from authors' Google Scholar profiles. This approach automates the labeling process, reducing cognitive effort and providing clearer insights into the topics. Experimental results demonstrate that the generated labels align effectively with the identified topics, confirming their suitability. Overall, the study makes significant strides in topic modeling by presenting innovative algorithms for improving coherence, diversity, and subject interpretation in scientific publications.

References

[1] Avasthi, S., Chauhan, R., & Acharjya, D. P. (2023). Extracting information and inferences from a large text corpus. International Journal of Information Technology, 15(1), 435-445.

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

[3] Shaker, N. H., & Dhannoon, B. N. (2024). Word embedding for detecting cyberbullying based on recurrent neural networks. Int J Artif Intell ISSN, 2252(8938), 8938.

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

[5] Mann, J. K. (2021). Semantic Topic Modeling and Trend Analysis.

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

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

[8] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112, 102131.

[9] Alemayehu, E., & Fang, Y. (2024). Supervised probabilistic latent semantic analysis with applications to controversy analysis of legislative bills. Intelligent Data Analysis, (Preprint), 1-23.

[10] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789-5829.

[11] Deepa, M. D. (2021). Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(7), 1708-1721.

[12] 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).

[13] Kozbagarov, O., Mussabayev, R., & Mladenovic, N. (2021). A new sentence-based interpretative topic modeling and automatic topic labeling. Symmetry, 13(5), 837.

[14] Avasthi, S., & Chauhan, R. (2024). Automatic label curation from large-scale text corpus. Engineering Research Express, 6(1), 015202.

[15] Helan, A., & Sultani, Z. N. (2023, February). Topic modeling methods for text data analysis: a review. In AIP Conference Proceedings (Vol. 2457, No. 1). AIP Publishing.

[16] Allahyari, M., Pouriyeh, S., Kochut, K., & Arabnia, H. R. (2017). A knowledge-based topic modeling approach for automatic topic labeling. International Journal of Advanced Computer Science and Applications, 8(9), 335.

[17] Kinariwala, S. A., & Deshmukh, S. (2021). Onto_TML: Auto-labeling of topic models. Journal of Integrated Science and Technology, 9(2), 85-91.

[18] Zhou, M., Kong, Y., & Lin, J. (2022, July). Financial Topic Modeling Based on the BERT-LDA Embedding. In 2022 IEEE 20th International Conference on Industrial Informatics (INDIN) (pp. 495-500). IEEE.

[19] Hristova, G., & Netov, N. (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) (pp. 2259-2264). IEEE.

[20] Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A study of text vectorization method combining topic model and transfer learning. Processes, 10(2), 350.

[21] George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. International Journal of Information Technology, 15(4), 2187-2195.

[22] Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. Applied Sciences, 13(2), 797.

[23] Talebpour, M., García Seco de Herrera, A., & Jameel, S. (2023). Topics in contextualised attention embeddings. In European Conference on Information Retrieval (pp. 221-238). Cham: Springer Nature Switzerland.

[24] Tang, G., Chen, X., Li, N., & Cui, J. (2023). Research on the evolution of journal topic mining based on the bert-LDA model. In SHS Web of Conferences (Vol. 152, p. 03012). EDP Sciences.

[25] 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.

[26] Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. arXiv preprint arXiv:1612.05340.

[27] Gourru, A., Velcin, J., Roche, M., Gravier, C., & Poncelet, P. (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 (pp. 352-363). Springer International Publishing.

[28] He, D., Ren, Y., Khattak, A. M., Liu, X., Tao, S., & Gao, W. (2021). Automatic topic labeling using graph-based pre-trained neural embedding. Neurocomputing, 463, 596-608.

[29] 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.

[30] Saeed, A., Khan, H. U., Shankar, A., Imran, T., Khan, D., Kamran, M., & Khan, M. A. (2023). Topic modeling based text classification regarding islamophobia using word embedding and transformers techniques. ACM Transactions on Asian and Low-Resource Language Information Processing.

[31] Salman, Z. A. W. (2023). Text Summarizing and Clustering Using Data Mining Technique. Al-Mustansiriyah Journal of Science, 34(1), 58-64.

[32] Gelar, T., & Sari, A. N. (2024, February). 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) (pp. 129-140). Atlantis Press.

[33] Navarro, E., & Homayouni, H. (2023). Topic Modeling in Cardiovascular Research Publications.

[34] Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084.

[35] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

IRAQI Academic Scientific Journals

Samarra Journal of Pure and Applied Science

SJPAS
مجلة سامراء للعلوم الصرفة والتطبيقية

www.sjpas.com

p ISSN: 2663-7405
e ISSN: 2789-6838

من البيانات إلى البصيرة: نمذجة المواضيع واستراتيجيات تصنيف المواضيع التلقائية

رنا فارس نجيب [1]*، بان نديم ذنون ، فرح قيس الخالدي[2]

1- قسم الحاسبات ، كلية العلوم ، الجامعة المستنصرية ، العراق

2- قسم الحاسوب ، كلية العلوم ، جامعة النهرين ، العراق

البحث مستل من رسالة الدكتوراه الباحث الاول

معلومات البحث:

تاريخ الاستلام: 2025/10/21
تاريخ التعديل: 2025/11/24
تاريخ القبـــول: 2025/11/28
تاريخ الـــنشر: 2025/12/30

الكلمات المفتاحية:

التعلم العميق، S-BERT، تقليل الأبعاد، تماسك الموضوع، وتنوع الموضوع

معلومات المؤلف

الايميل:

الخلاصة:

من أجل تعزيز إمكانية تفسير البيانات لأغراض صنع القرار، تتطلب مجموعات النصوص العلمية والبيولوجية والاجتماعية تقنيات التعلم الآلي الفعالة. يتم دعم التنقيب عن النصوص من خلال نماذج المواضيع في مصادر مثل المدونات، وبيانات تويتر، والمجلات العلمية، والأوراق الطبية الحيوية. لا يزال من الصعب العثور على التصنيفات المناسبة، حتى عندما تشير نماذج المواضيع إلى مفاهيم مهمة. يتم تقليل الجهد المعرفي للمحللين عن طريق أتمتة تقييم الموضوع وتصنيفه. في حين أن بعض التقنيات تعتمد على تكرار الكلمات لإنتاج تسميات تحتوي على كلمات أو عبارات أو صور، فإن الطرق الاستخراجية تختار التسميات بناءً على مقاييس الاحتمالية. تقترح هذه الدراسة تحسين نمذجة المواضيع في مجموعة من أوراق المؤتمرات حول أنظمة معالجة المعلومات العصبية (NIPS) التي تم إصدارها بين عامي 1987 و2017 وحققت هدفين: إنتاج موضوعات أكثر تماسكًا ووضع العلامات التلقائية على المواضيع. تم تحقيق الهدف الأول من خلال خمس مراحل: مرحلة المعالجة المسبقة للنص، مرحلة التخفيض باستخدام طريقة جديدة تسمى SR-LW (تقليل الجمل على أساس الطول والوزن)، والتي تزيل الجمل الأقصر طولًا، ثم تحسب وزن الجمل المتبقية. ويزيل ما يقرب من 25% من الجمل ذات الوزن الأقل. تستخدم مرحلة تضمين الجملة S-BERT (تمثيل تشفير الجملة ثنائي الاتجاه من المحول) لتقليل أبعاد مرحلة تضمين الجملة من خلال استخدام التقريب والإسقاط المتنوع الموحد (UMAP). وأخيرًا، نظم التجميع المكاني الهرمي القائم على الكثافة للتطبيقات ذات الضوضاء (HDBSCAN) وثائق قابلة للمقارنة. توضح النتائج التجريبية أن استخدام مرحلة SR-LW المقترحة قد أنتج موضوعات أكثر تماسكًا، مما أدى إلى تحسين تماسك الموضوع بمقدار (0.593) وأداء تنوع الموضوع بمقدار (0.96). على الرغم من أن نمذجة الموضوع تستخرج الجمل الأكثر بروزًا التي تصف الموضوعات الكامنة من المجموعات النصية، إلا أنه لم يتم تحديد التسمية المناسبة بعد. أما الهدف الثاني فقد تم تحقيقه من خلال اقتراح طريقة جديدة لتوليد الكلمات الرئيسية من خلال الوصول إلى الملفات الشخصية للمؤلفين في الباحث العلمي من Google واستخراج الاهتمامات لاستخدامها في تصنيف المواضيع تلقائيًا.