UDK:
004.6

10.58898/sij.v2i2.01-13

TEXT MINING: CLUSTERING USING BERT AND PROBABILISTIC TOPIC MODELING

Kavitha, D.[1], Anandha Mala, G. S.[1], Padmavathi, B.[1], Varshni. S. V.[1]*

[1]Easwari Engineering College, Ramapuram, Chennai, 600089, Tamil Nadu, India.
e-mail: dkavithavijay@gmail.com, gs.anandhamala@gmail.com, bpadma_cse@yahoo.com, varshnisv2021@gmail.com

Abstract: In order to find significant patterns and fresh ideas, free-form content is transformed into structured format using a process known as text mining or text data mining. It enables businesses to easily locate important information in texts like emails, social media posts, support requests, chatbots, and other sorts of text. Text mining enables businesses to anticipate possible threats from rivals, react quickly to production or delivery problems, and provide more individualised customer service. Businesses employ text mining for a range of functions, including production, IT, marketing, sales, and customer service. By carefully examining the phrases used in the source texts, topic modelling aims to pinpoint the recurrent themes in a corpus. These concepts are known as “topics”. As a result, textual data may be measured and used in quantitative analysis. In this sector, there are several subject modelling kinds that differ from one another based on a few unique traits and criteria. In our paper we have represented mainly 3 types of topic modelling techniques namely Latent Semantic Analysis (LSA), Hierarchical Dirichlet Process (HDP), and Latent Dirichlet Analysis (LDA) and calculated the coherence score of each method and compared them. And we have infused the concept of BERT with this topic modelling models and proposed a new model called HDP BERT and calculated the coherence Score and clusters the topics. At the end the n-grams features are applied to all 4 models and compared among each other in bases of uni, bi and trigram rate percentage.

Keywords: LDA, LSA, BERT, HDP, N-grams.

1.0 INTRODUCTION

The technique of converting unstructured material into a structured format is known as text mining, also known as data mining for text, and it is utilised to find significant patterns and fresh perspectives. Companies are able to investigate and find by utilising cutting-edge such as Support Vector Machines (SVM), Naive Bayes, and other deep learning algorithms are examples of analytical methods, they may uncover hidden correlations within their unstructured data. Although in everyday speech the phrases generally speaking, text mining and text analytics can also indicate different things. Text mining and text analysis use a combination to discover textual patterns and trends in unstructured data using machine learning, statistics, and linguistics. By organising the data via means of text analysis and mining, text analytics may be used to discover quantitative insights. Afterwards, you may use data visualization tools to share your results with more people .In the process of pre-training deep bidirectional representations from unstructured texts, each left and right meaning are supposed to be synchronously conditional. Adding only one more output layer, for a variety of NLP applications, the pre-trained BERT model may be enhanced to provide cutting-edge models. [17] BERT succeeds in this from before the phase, which accounts for 50% of its success. This is due to the fact that a model learns to understand the text more thoroughly and accurately while being trained on a large corpus. This information is like a Swiss army knife that can be applied to practically any NLP assignment.

On the contrary, N-grams are a text’s repeated word, symbol, or character sequences. Technically, they may be referred to as the neighbouring groups of objects in a document. These were important when text data is processed using natural language. At the same time in machine learning, we typically group instances as a first step in understanding a subject (data set) in a machine learning system. The process of gathering unlabelled samples is called clustering. Machine learning without supervision is used for clustering since the examples are unlabelled. Labelling the instances makes clustering become categorization. See Introduction to Machine Learning Problem Framing for a more in-depth explanation of

[*] Corresponding author: dkavithavijay@gmail.com

supervised vs unsupervised approaches. Topic Modelling is a technique used by genetic search engines to gather and deliver pertinent content in response to user queries. The methods used to organize and portray information are more important than the Topic Modelling application, despite the fact that it appears straightforward.

A common method for locating underlying themes in a group of texts is topic modelling, which is as a tool for natural language processing and machine learning. Most of the top subject modelling ideas are shown below: Latent Dirichlet Allocation (LDA): The foundation of this well-known generative probabilistic model is the notion that each text is made up of a variety of themes, and that each theme is a word-by-word probability distribution. By applying the non-negative matrix factorization (NMF) method, a matrix is split into two non-negative matrices. Each document is represented by NMF as a linear combination of topic vectors for topic modelling. Hierarchical Dirichlet Process (HDP): This is a Bayesian non-parametric technique that models each document as a combination of an unlimited number of themes. HDP can determine the underlying structure of subjects and subtopics in a collection of texts. Correlated Topic Model (CTM): This is a continuation of LDA that models the correlation between topics. CTM implies that topics are created by a multivariate Gaussian distribution, which allows for the modelling of topic correlation. Probabilistic Latent Semantic Analysis (PLSA): This is a model that generates ideas where each subject is assumed to be a probability distribution over words and each text to be a mixture of themes. PLSA is similar to LDA, except it does not make the assumption of a Dirichlet prior on the topic distribution. Natural language processing tasks have been taught to the Transformers BERT - Bidirectional Encoder Representations deep learning model including interpreting language and responding to questions. BERT [21,24,25], on the other hand, may be utilized for topic modelling by employing the idea of fine-tuning.

Although prominent topic modelling approaches Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) have inherent limitations. The sensitivity of LDA to the number of subjects (K) parameter is one of its drawbacks. Finding the right value for K might be difficult. The model could oversimplify and combine different themes into one if K is set too low. However, if K is set too high, the model could become very complex and produce erroneous subjects. Additionally, LDA bases its analysis on the assumption that each document contains a variety of themes, which may not necessarily be the case for every dataset, yielding less precise findings. LSA, however, has drawbacks since it relies heavily on linear algebra and the singular value decomposition (SVD) method. Due to processing limitations, managing very big datasets is frequently difficult. Additionally, word order is not taken into account by LSA, which might cause context and subtlety to be lost in the text. It considers each word separately, which is a simplification that could not accurately reflect the nuances of genuine language.

Both LDA and LSA have the “bag of words” issue, where they fail to take into account the texts’ structural and sequential information. It may be difficult to simulate subjects that depend on word choice or context because of this restriction. Both approaches may also have trouble with polysemous words (words with numerous meanings) and may not be able to clearly separate out various uses of the same term.

In comparison to conventional topic modelling methods like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), using the Hierarchical Dirichlet Process (HDP) in conjunction with BERT (Bidirectional Encoder Representations from Transformers) has a number of benefits. First off, HDP with BERT offers greater topical freedom when modelling. In contrast to LDA and LSA, which demand that the number of topics be pre-specified, HDP enables an infinite number of topics to arise from the data. When working with huge, diversified datasets where the number of underlying themes is unknown or expected to fluctuate over time, this flexibility is very helpful. It guarantees that a wide range of complex and changing themes in the text may be captured by the model.

Second, the integration of HDP and BERT benefits from the contextualised word embedding of BERT. As opposed to the conventional bag-of-words method employed by LDA and LSA, BERT captures complex semantic links between words by taking into account their context in a phrase. As a result, topic modelling findings are more precise and contextually sensitive, enabling the model to distinguish minute variations in context and meaning across words and sentences. Additionally, HDP with BERT is ideally suited for addressing polysemy and homonymy, problems that LDA and LSA struggle with. The model can distinguish between words with different meanings based on their context in a document thanks to the contextual embedding in BERT, which leads to a more accurate representation of subjects.

2.0 RELATED WORK

A semantic content analysis topic modelling objective was achieved in [1] by utilising the N-gram model and the Latent Dirichlet Allocation (LDA) topic modelling approach on 3562 peer-reviewed scientific publications Since the commencement of the COVID-19 outbreak, there have been several publications. The study's findings are expected to be useful for researchers and future research. Similarly, other studies [2], they introduce a topic-based adversarial learning model for finding videos with bogus news. However, it would be ineffective to use traditional topic models directly to such brief texts. [3] The main reasons why the current research there is a paucity of such work for Urdu text and discovered niche themes even using conventional topic models are a lack of benchmark datasets, due to time and computational constraints, a lack of benchmark datasets, a lack of pre-processing tools and methods, and limitations on large-scale datasets. The NMF, PLSA, LSA, and LDA topic modelling techniques were tested on 0.8 million Urdu tweets in this study. To prevent the dataset from being overly focused on a specific topic, a collection of tweets was obtained using the Twitter API, utilizing a variety of hashtags as search queries. Additionally, the dataset was subjected to three distinct variations, underwent text pre-processing to standardize the tweet content, and underwent a process to represent the tweets as documents using various n-grams. They have also shown how these methods work using visualisation techniques, graphs showing the number of tweets per subject, word clouds, and hashtag analysis, which provide information about how well the algorithms perform on certain themes.

NMF fared better utilising the TF-IDF feature set while integrating short-text method into large pseudo documents. According to the results, NMF performed better when employing LDA performed better while analysing TF-IDF feature vectors in the text of Urdu tweets when short-text method and long pseudo documents were combined. [4] This study revisits LDA-style methods and uses a theoretical framework to examine how word co-occurrence and topic models relate to one another and their investigation demonstrates that topic discovery may be improved by changing the word co-occurrences within the corpus. As a result, they suggest a cutting-edge data modification technique called DATM to enhance topic discovery inside a corpus. In [5,6] The review text that crowdsourcing participants get from employers contains vital knowledge, views, and preferences that can be used by both parties to enhance the level and quality of their services. According to experimental findings, adding the classifier now has LDA topic model characteristics and finds it is difficult to differentiate the sentiment categories of multiple emotion polarity terms that coexist in text, but review text may effectively handle this issue and improve the performance of emotion border fuzzy text categorization. With an accuracy of 0.881, the GBDT text sentiment classifier performs the best in terms of classification; the F1-measures of the samples from the second, third, fourth, and fifth categories are, correspondingly., 0.462, 0.571, 0.278, and 0.647. It also has applications in the field of The results of using this general method in the field of information security provide evidence for its viability.[7] The collection of substantial text data regarding the Anti-ELAB Movement from a well-known forum in Hong Kong is used to construct a complex BERT-LDA hybrid model for large-scale network public opinion analysis. They also look at the emotional makeup and development of public opinion in connection to the "text topic," as well as the characteristics and positions of opinion leaders during Anti-ELAB public opinion campaigns. [8] TECM-JD which is The topic-extended emotional conversation generation model based on joint decoding is proposed in this study. The experiment's findings demonstrate that the suggested model generates richer emotional content that is relevant to the subject and functions effectively and outperforming traditional conversation models. However, while creating dialogue, we frequently overlook the importance of the subject and context information. As a result, the produced replies are usually devoid of context or result in wide responses due to a lack of topic knowledge. In this study, [9], the authors look at how multi-turn conversations grow using a large corpus, and they make use of the discussion's subject and context information throughout the dialogue generation process to provide more persuasive, context-sensitive responses. By employing this approach, the model can ingest more contextual information, enhancing its ability to generate high-quality responses that are contextually relevant. Consequently, even in the case of languages with abundant resources, uncovering meaningful latent topics from online chat texts can be challenging, particularly when dealing with low word co-occurrence patterns and limited access to extensive social media benchmark datasets. The absence of such benchmark datasets, coupled with a scarcity of pre-processing tools/algorithms, constraints in terms of time and computational resources when dealing with large datasets, and the overall scarcity of benchmark datasets, collectively contribute to the existing research gap. This gap exists in the context of Urdu text analysis, where discovering specialized topics remains a challenge, even when employing traditional topic modelling techniques. Additionally, they have shown the effectiveness of these strategies using visualisation techniques, graphs showing the number of tweets per subject, word clouds, and hashtag analysis, providing information on how well the algorithms perform on selected themes.

3.0 PROPOSED METHOD

In below figure 4 it is represented that how the overall the proposed system works, where the dataset CORD 19 is provided as an algorithm input in which the provided dataset is pre-processed and data cleaning takes place and further the required data is extracted and BERT features are applied along with the k-means clustering.

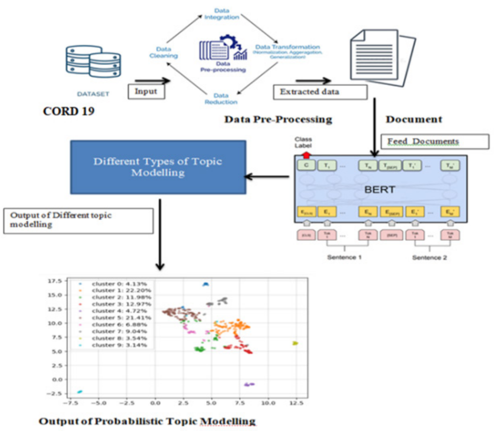


Figure 1. System Architecture of HDP_BERT

3.1 Data Collection

Data collection is the act of obtaining and examining precise data from a variety of sources to identify solutions to study issues, trends, probabilities, etc., and to assess potential consequences. The dataset CORD 19, which is used in this work, was used to gather the data. More than a million research articles regarding COVID-19, SARS-CoV-2, and similar coronaviruses, with over 400,000 of them being full-text. may be found in the CORD-19 database. The research community as a whole may use this openly available data to offer new insights that will help in the ongoing fight against this infectious sickness with the aid of recent advancements in NLP and other AI Methods. The medical research community is struggling to keep up with the tremendous acceleration of new coronavirus literature, which has increased the necessity for these techniques. After collecting the relevant dataset (CORD 19) the required set was journal papers should be extracted, so we give a condition that the papers should be visible only after Jan 2020 and proceed by extracting the abstract of all this papers only with a condition of set of papers with indexing of each of them in order for easy recognition of each paper by using index.

3.2 Data Pre-Processing

Data preparation is a crucial step in the data mining process. It explains the steps involved in cleaning, transforming, and combining data in order to prepare it for analysis. Data preprocessing is done to raise the standard and applicability of the data for the specific data mining activity. In our proposal first the sentence level pre –process takes place where The way the sentences are put together The grammatical rules of the language are primarily what control NLP. We speak of phrase structure grammar when referring to sentence building. A few examples of phase structure grammar are Grammars with generalized phrase structure, lexical functional grammar, phrase structure driven by the head, etc. The construction of these grammars consists mostly of several feature structures, which are nothing more than a collection of attribute-value pairs. Further added to it the Normalization takes place for the following like Missing delimiters, lower cases, letter repetition, non-word repetition and etc. Language detection takes place where for our model we have given English, French, Spanish and Chinese. After Sentence level pre-process the word pre-process takes place where the filtering of Punctuations and numbers takes place and selecting nouns, typo correction and filtering of Stop words list is done.

An examination of the content was carried out utilizing a word-level N-gram model to reveal the distinct semantic structures and terminology commonly utilized in the corpus. The N-gram model classifies linguistic units into the following categories: unigrams representing individual words, bigrams representing pairs of two words, trigrams representing groups of three words, and so forth. This analysis entailed the measurement of the frequency of occurrence of unigrams, bigrams, and trigrams within the corpus. This approach aimed to identify and quantify domain-specific language structures and terminology commonly used in the dataset. It was possible to identify the corpus's key semantic structures and terminologies.

3.3 Auto Encoder

- Probabilistic topic assignment vector using LDA.
- Bert for the vector of a sentence.

A weight hyper parameter was used to concatenate the LDA and Bert vectors [15] in order to balance the relative weights of the data from each source. A lower dimensional latent space representation of the concatenated vector was learned using an auto encoder. The concatenate vector is assumed to have a manifold space in the high-dimensional space. Used clustering to identify themes using latent space representations. Architecture of the auto encoder for learning latent space representation is condensed to only one hidden layer.

3.4 Fitting The Models with BERT

In our proposal we are implementing 3 types of existing models along with a newly proposed model namely:

- LDA-BERT
- LSA-BERT
- HDP-BERT (Proposed model)

Basically in this module the different types of models are combined with BERT concept and the coherence Score is calculated for each model and compared among them. To find ''themes,'' or latent semantic patterns in a body of text, An approach to machine learning is topic modeling which is frequently used in NLP and text mining. Topic modeling has a foundation in probabilistic methods and statistical algorithms. For semantic content analysis, the probabilistic and generative LDA (Latent Dirichlet Allocation) algorithm, which employs a methodical approach, is a useful subject in modeling methodology. Similar way LSA is Latent Semantic Analysis (LSA) can play a part in this situation. In order to identify the hidden or latent concepts, also known as themes, LSA seeks to take use of the context around the words. So it won't be very beneficial for us if we just map the words to documents. The underlying ideas or subjects that lie beneath the words are what we truly need to find. And A Potent in the realm of unsupervised analysis for clustered data, the hierarchical Dirichlet process (HDP) stands as a mixed-membership model. Unlike its finite counterpart, latent Dirichlet allocation, the HDP topic model deduces the number of topics directly from the data, rather than relying on a predefined number of topics. In this module the vectorization and tokenization of the words takes place based on the different working model of the different models of topic modeling. The BERT is used for embedding in this process.

3.5 Clustering using K-Means

The unsupervised learning algorithm, K-Means, partitions an unlabeled dataset into distinct categories using clustering techniques. K-Means requires a predetermined number, K, of clusters to be established during the process. For instance, if K=2, the data is divided into two clusters; if K=3, three clusters are created, and so on. This centroid-based method assigns a centroid to each cluster and aims to minimize the distances between data points and their associated clusters. Initially, the algorithm separates the unlabeled input dataset into K clusters, iterating until no more clusters can be formed. In our model, we applied this clustering approach to all four models, represented the results using word clouds, and calculated the silhouette score for each method, facilitating a comparison of their effectiveness.

3.6 Proposed Algorithm HDP - BERT

```
Step 1: Dataset Preparation
# Download and preprocess the CORD-19 dataset
# Assume you have a list of preprocessed documents: pre-processed_docs

Step 2: Tokenization and Encoding
# Use BERT tokenizer to tokenize the preprocessed text into sub word tokens
# Convert tokens into BERT input features
```

```
Step 3: BERT Encoding
# Encode input features using a pre-trained BERT model
# Extract contextualized word embedding or sentence embedding

Step 4: HDP Topic Modeling
# Apply HDP on the BERT embedding

Step 5: K-means Clustering
# Apply K-means clustering on the HDP representation
# Determine the optimal number of clusters using techniques like the elbow method or silhouette score

Step 6: Silhouette Score Calculation
# Calculate the Silhouette score to evaluate the clustering performance

Step 7: Coherence Score Calculation
# Calculate the coherence score to evaluate the topic model quality
```

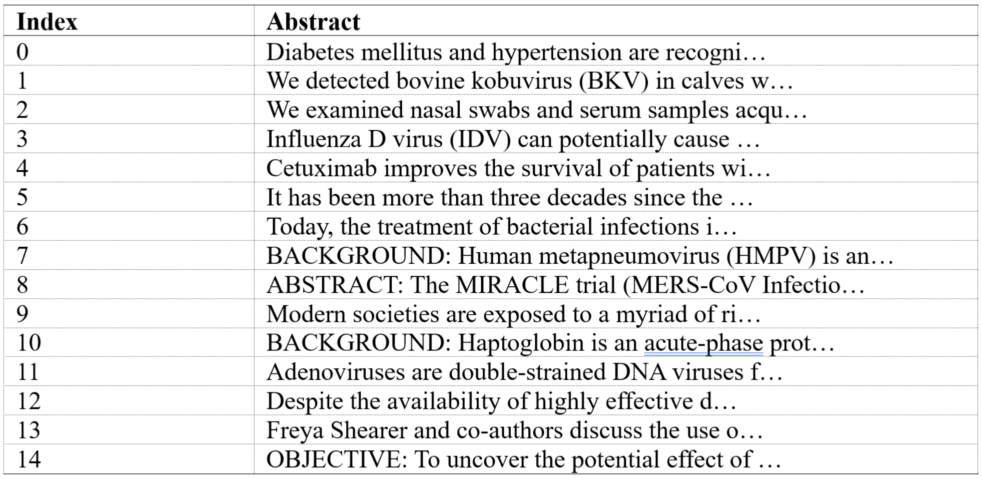
The preparation and BERT encoding [12,27,23] of the dataset are described in stages 1 and 2 of the aforementioned algorithm. In step three, contextualized word embedding or sentence embedding are obtained once the input features are encoded using BERT. The BERT embedding are subjected to HDP topic modelling in step four. The HDP representation is clustered in Step 5 using K-means clustering. In steps 6 and 7, the performance of the clustering is assessed using the Silhouette score, and the topic model's quality is assessed using the coherence score.

4.0 PERFORMANCE ANALYSIS

4.1 Descriptive Analysis

To gain an understanding of the descriptive analysis, a total of 51078 articles were analysed, with the count reduced to 5330 after applying the condition that the articles needed are for after January 1, 2020, and the count further reduced to 3947 of the articles and their abstracts are displayed with index to make it easy to identify the abstract of each article. And their output with abstract is included in Table 1.

Table 1. Abstract extraction of the articles which were published after JAN 1 2020 from CORD 19 dataset, top 15 papers were displayed with index.



Index	Abstract
0	Diabetes mellitus and hypertension are recogni…
1	We detected bovine kobuvirus (BKV) in calves w…
2	We examined nasal swabs and serum samples acqu…
3	Influenza D virus (IDV) can potentially cause …
4	Cetuximab improves the survival of patients wi…
5	It has been more than three decades since the …
6	Today, the treatment of bacterial infections i…
7	BACKGROUND: Human metapneumovirus (HMPV) is an…
8	ABSTRACT: The MIRACLE trial (MERS-CoV Infectio…
9	Modern societies are exposed to a myriad of ri…
10	BACKGROUND: Haptoglobin is an acute-phase prot…
11	Adenoviruses are double-strained DNA viruses f…
12	Despite the availability of highly effective d…
13	Freya Shearer and co-authors discuss the use o…
14	OBJECTIVE: To uncover the potential effect of …

As seen in table 1 it is clear with the output of Index assigned for each paper which as abstract and the papers are of after 1 JAN 2020. The index assigned for each paper is unique and Top 15 papers are displayed.

4.2 Coherence Score

In this section the Coherence Score calculated for all 4 models is discussed and compared among them in order to know which model is best in coherence score. By calculating the level of semantic resemblance between the subject's top phrases and the subject A single subject is evaluated for coherence. These metrics help distinguish between topics that are artefacts of statistical inference and problems that may be interpreted semantically. The coherence score in topic modeling [10] may be used to gauge how comprehensible the subjects are to people. Topics are shown below as the top N words that are most likely to fall under a particular category. The overall similarity of these words to one another is measured by the coherence score. The more the coherence score the best the model is. In this section we will calculate 4 types of model coherence score namely,

- LDA-BERT
- LSA-BERT
- HDP
- HDP-BERT

Let:
"D" represents the total number of documents in the corpus.
"C (A, B)" represent the co-occurrence count of words A and B in the corpus.
"C(A)" represent the occurrence count of word A in the corpus.
Then, the U-Mass coherence score formula using alphabet letters can be written as:

$$U\_mass = \Sigma (\log ((C (A, B) + 1) / C(A)))$$

This formula calculates the U-Mass coherence score for a set of word pairs within a topic. You would calculate the coherence score (log ((C (A, B) + 1) / C(A))) for all word pairs in the topic and then sum up these values to obtain the coherence score for that topic

The above is the formula used to calculate coherence Score in HDP BERT.

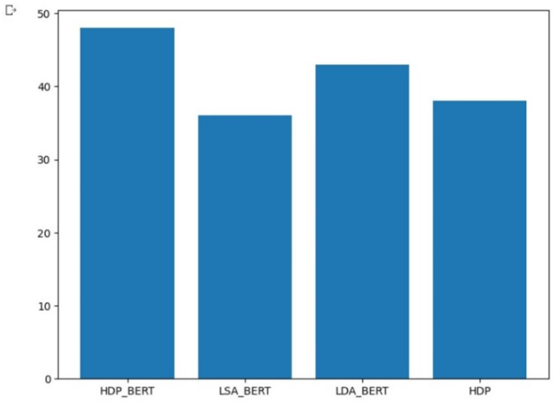


Figure 2. Comparison of 4 models with respect to Coherence Score

In figure 2 all the 3 existing model (LDA with BERT, LSA with BERT and HDP) is compared with the proposed model HDP BERT. From above figure it is clear that the HDP_BERT occupies first position with the coherence score of 0.48, next LDA with BERT 0.42, next HDP with 0.38 score and at the end it is LSA BERT of 0.36.

Table 2. Tabular comparison of all 4 models with respect to Coherence Score

Models	Coherence Score
HDP_BERT	0.49223918030943364
LDA_BERT	0.44805456363544180
HDP	0.37751452573322253
LSA_BERT	0.37212732811287774

4.3 Silhouette Score

A statistic used to assess the effectiveness of the clustering process is the silhouette score. A The compactness of each cluster (intra cluster distance) and the distance between clusters (inter cluster distance) are used to compute the overall representative score of the efficacy of our clustering approach. The value of the silhouette coefficient is between [-1, 1]. The best score is 1, which shows that the data point i is far from other clusters and is very compact inside the cluster to which it belongs. -1 is the poorest possible value. Through our observation it is was analyzed the following reading of Silhouette score for all 4 models. Below is the formula used to calculate the Silhouette Score.

Let:
"a_i" be denoted as "a of i"
"b_i" be denoted as "b of i"
"N" represents the total number of data points (samples)
"N_s" represent the number of data points in cluster "s"
"s(i)" represent the silhouette score for data point "i"
Then, the Silhouette Score formula with the new variable names can be written as:
s(i) = max (a of i, b of i) / max (b of i, a of i)
Silhouette Score = (1 / N) * Σ (N_s * s(i)) for i = 1 to N

Table 3. Tabular comparison of all 4 models with respect to Silhouette score

Models	Silhouette Score
HDP_BERT	0.39303713
LDA_BERT	0.39769357
HDP	clustering was not able to be executed
LSA_BERT	0

In above mentioned Table 3 it is clear that for LDA BERT the score is more compared to the score of HDP BERT but they are almost same only in few point values they vary therefore we can come to an conclusion that these two methods are best in clustering whereas apart from these two methods there are other 2 methods namely LSA BERT and Normal HDP. For both of these methods the LSA BERT has the score of 0 which is the worst case in Silhouette score whereas it was not able to perform the clustering operation because the clustering operation can be done only by combining BERT concept to the model then only regression takes place.

4.4 Clustering and Word Cloud

Unsupervised learning is a technique. In machine learning or data science, clustering issues are addressed with K-Means Clustering. The K-means clustering technique is described in this chapter, along with instructions on how to use Python to carry it out [11]. Without the need for any training, it provides a useful method for automatically categorizing the groups in the unlabeled dataset. It may also be used to group the data into several clusters. This centroid-based method connects each cluster to a centroid. Reducing the overall distances between the data points and the clusters they are a part of is the main objective of this strategy. The program initially takes as input an unlabeled dataset and divides it into k clusters. The program then keeps doing this until there are no more clusters available for use. In this procedure, the value of k must be preset. This group of words is shown in various sizes and is also known as a word cloud. The more frequently and critically a phrase was used, the larger and bolder it will appear in the document. These techniques, sometimes known as tag clouds or text clouds, are effective for extracting the most significant parts of textual information, including blog posts and databases. They can also help business users find phrase overlaps by comparing and contrasting two separate documents.

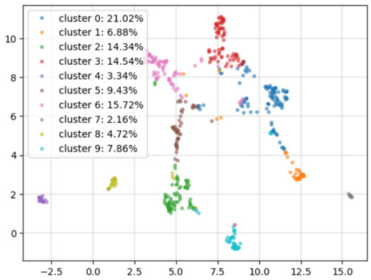


Figure 3. Clustering of LDA BERT using K-means

In figure 3 it is shown the clustering of LDA BERT using K-means. Here the clustering is given as 10 count therefore the number of cluster in above figure is 10 from 0 to 9. Each cluster carries unique colors and this colors are actually used in the topic word representation in word cloud. It is displayed like the cluster 0 is in color of light blue with 21.02% of topic mining, followed by cluster 1 with 6.88% of orange color similarly different clusters are of different colors.

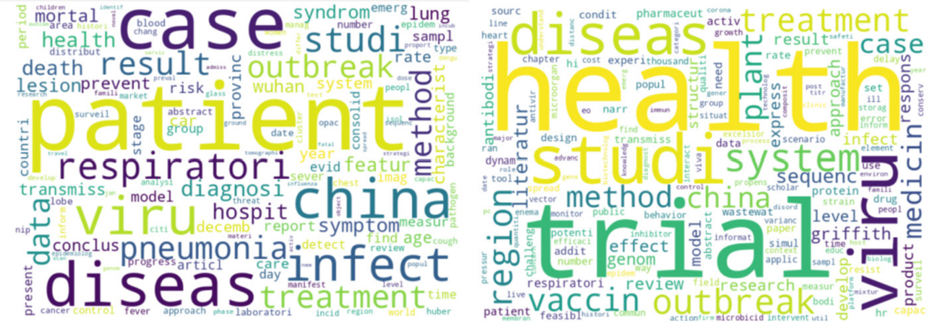


Figure 4. Word cloud representation of LDA BERT

From above figure 4 it is represented the word cloud look of the clustering using K-means done on the LDA _BERT model. Actually the number of clustering done on the model is 10 but for a sample only 2 clustering is shown. The words like 'case', 'patient', 'china' 'diseas', 'pneumonia' and 'infect' are the most repeated words or topics from the cluster 0, whereas the words which are small are the words which are repeated for less count example 'death', 'health', 'moral', 'number' and etc. From the cluster 1 also we can find the same concept and pattern.

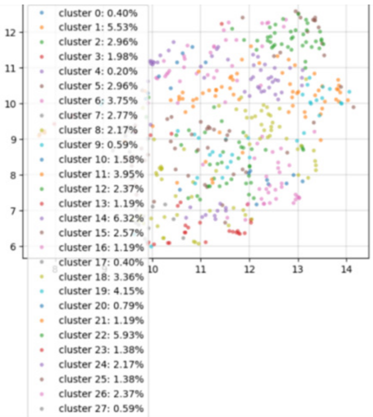


Figure 5. Clustering of HDP BERT using K-means

In figure 5 it is shown the clustering of HDP BERT using K-means. Here the clustering is given as 10 count therefore the number of cluster in above figure is 50 from 0 to 49. Each cluster carries unique colors and this colors are actually used in the topic word representation in word cloud. It is displayed like the cluster 0 is in color of light blue with 0.40 % of topic mining, followed by cluster 1 with 5.53% of orange color similarly different clusters are of different colors. Compare to LDA_BERT the clusters are more scattered because in this model the number of cluster count is given is more because with more iteration and number of cluster more also the coherence score is high this shows that HDP_BERT model is best.

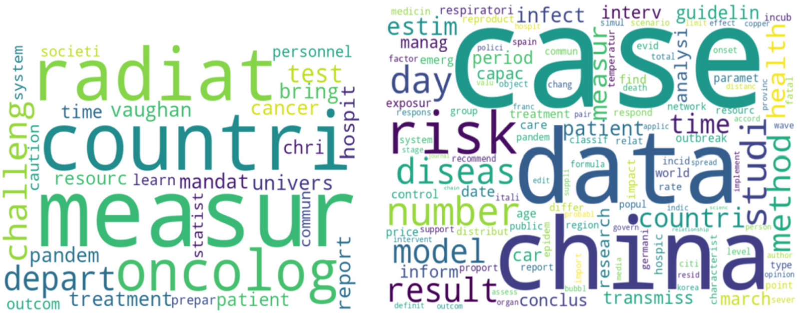


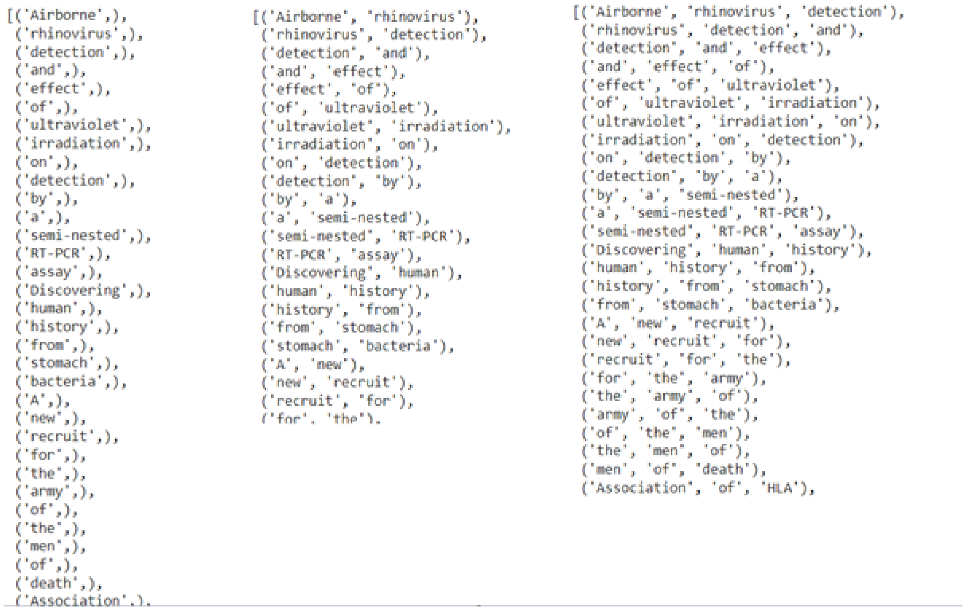
Figure 6. Word cloud representation of HDP BERT

From above figure 6 it is represented the word cloud look of the clustering using K-means done on the HDP _BERT model. Actually the number of clustering done on the model is 50 but for a sample only 2 clustering is shown. The words like 'measure', 'depart', 'countri', 'challenge' and etc. are the most repeated words or topics from the cluster 0, whereas the words which are small are the words which are repeated for less count example 'outcome', 'patient', 'learn', 'cancer' and etc. From the cluster 1 also we can find the same concept and pattern.

4.5 N-Grams

A text document that has n consecutive objects, including words, numbers, symbols, and punctuation, is known as an n-gram [14,16,18]. In many text analytics applications where word sequences are important, such as sentiment analysis, text categorization, and text production, N-gram models are helpful. N-grams are ongoing groups of letters, numbers, or other symbols that appear in a document. Technically speaking, they are the adjacent groups of items in a document. They are relevant when performing NLP (Natural Language Processing) activities on text data. They can be used for a variety of things, including text mining, machine translation, spelling correction, semantic characteristics, language models, and semantic features. If only one word if we want it is called unigram, when it is 2 words it is bi gram, and for 3 words it is trigram.

Table 4: unigram(left), bigram(center), trigram(right)



```
[('Airborne',),
 ('rhinovirus',),
 ('detection',),
 ('and',),
 ('effect',),
 ('of',),
 ('ultraviolet',),
 ('irradiation',),
 ('on',),
 ('detection',),
 ('by',),
 ('a',),
 ('semi-nested',),
 ('RT-PCR',),
 ('assay',),
 ('Discovering',),
 ('human',),
 ('history',),
 ('from',),
 ('stomach',),
 ('bacteria',),
 ('A',),
 ('new',),
 ('recruit',),
 ('for',),
 ('the',),
 ('army',),
 ('of',),
 ('the',),
 ('men',),
 ('of',),
 ('death',),
 ('Association',),
```

```
[('Airborne', 'rhinovirus'),
 ('rhinovirus', 'detection'),
 ('detection', 'and'),
 ('and', 'effect'),
 ('effect', 'of'),
 ('of', 'ultraviolet'),
 ('ultraviolet', 'irradiation'),
 ('irradiation', 'on'),
 ('on', 'detection'),
 ('detection', 'by'),
 ('by', 'a'),
 ('a', 'semi-nested'),
 ('semi-nested', 'RT-PCR'),
 ('RT-PCR', 'assay'),
 ('Discovering', 'human'),
 ('human', 'history'),
 ('history', 'from'),
 ('from', 'stomach'),
 ('stomach', 'bacteria'),
 ('A', 'new'),
 ('new', 'recruit'),
 ('recruit', 'for'),
 ('for', 'the'),
```

```
[('Airborne', 'rhinovirus', 'detection'),
 ('rhinovirus', 'detection', 'and'),
 ('detection', 'and', 'effect'),
 ('and', 'effect', 'of'),
 ('effect', 'of', 'ultraviolet'),
 ('of', 'ultraviolet', 'irradiation'),
 ('ultraviolet', 'irradiation', 'on'),
 ('irradiation', 'on', 'detection'),
 ('on', 'detection', 'by'),
 ('detection', 'by', 'a'),
 ('by', 'a', 'semi-nested'),
 ('a', 'semi-nested', 'RT-PCR'),
 ('semi-nested', 'RT-PCR', 'assay'),
 ('Discovering', 'human', 'history'),
 ('human', 'history', 'from'),
 ('history', 'from', 'stomach'),
 ('from', 'stomach', 'bacteria'),
 ('A', 'new', 'recruit'),
 ('new', 'recruit', 'for'),
 ('recruit', 'for', 'the'),
 ('for', 'the', 'army'),
 ('the', 'army', 'of'),
 ('army', 'of', 'the'),
 ('of', 'the', 'men'),
 ('the', 'men', 'of'),
 ('men', 'of', 'death'),
 ('Association', 'of', 'HLA'),
```

In table 4 it is represented the unigram, bigram and trigram result after applying the n-gram features to dataset (CORD 19) [1]. It is clear from the above figure that unigram consists of only 1-word example 'Airborne', 'rhinovirus, detection' and etc. whereas in Bi-gram the number of words is of count 2 example 'Airborne and Rhinovirus' are taken or read together followed by in Tri-gram the count of words is of 3 words and here the number of words are read together is 3 example 'Airborne, rhinovirus, detection' is read together.

5.0 RESULT AND DISCUSSION

This paper aimed and proposed the best topic modelling method called HDP-BERT with respect to the coherence Score of the model. By doing this analysis we can predict the current trends and research interests in any particular fields. As in our proposal we used the CORD 19 dataset [1] which is actually a dataset. The CORD-19[1] database contains more than a million research publications regarding COVID-19, SARS-CoV-2, and similar coronaviruses, including more than 400,000 full-text articles. With the help of current developments in natural language processing and other AI approaches, the research community at large may use this freely accessible information to provide fresh insights that will aid in the ongoing battle against this contagious illness. The medical research community is struggling to keep up with the tremendous acceleration of new coronavirus literature, which has increased the necessity for these techniques. After collecting the relevant dataset (CORD 19) the required set was journal papers should be extracted, so we give a condition that the papers should be visible only after Jan 2020 and proceed by extracting the abstract of all this papers only with a condition of set of papers with indexing of each of them in order for easy recognition of each paper by using index.

To gain an understanding of the descriptive analysis, a total of 51078 articles were analysed, with the count reduced to 5330 after applying the condition that the articles needed are for after January 1, 2020, and the count further reduced to 3947 of the articles and their abstracts are displayed with index to make it easy to identify the abstract of each article.

The coherence score in topic modelling may be used to gauge how comprehensible the subjects are to people. Topics are shown below as the top N words that have the highest likelihood of falling under a certain subject. The coherence score evaluates how similar these words are to one another in general. The more the coherence score the best the model is. From Table 2 it is clear that the HDP_BERT occupies first position with the coherence score of 0.48, next LDA with BERT 0.42, next HDP with 0.38 score and at the end it is LSA BERT of 0.36.

The silhouette coefficient's value ranges from [-1, 1]. The best score is 1, which indicates that the data point i is highly compact inside the cluster to which it belongs and is located distant from other clusters. A value of -1 is the worst. Through our observation it is was analyzed the following reading of Silhouette score for all 4 models. In above mentioned Table 3 it is clear that for LDA BERT the score is more compared to the score of HDP BERT but they are almost same only in few point values they vary therefore we can come to an conclusion that these two methods are best in clustering whereas apart from these two methods there are other 2 methods namely LSA BERT and Normal HDP. For both of these methods the LSA BERT has the score of 0 which is the worst case in Silhouette score whereas it was not able to perform the clustering operation because the clustering operation can be done only by combining BERT concept to the model then only regression takes place.

6.0 CONCLUSION AND FUTURE WORK

The aim of this research was to delineate the research context described by these inquiries. Within this study, supported by three fundamental components, the initial step involved conducting descriptive analyses to identify the bibliometric attributes of the subject. Following the application of K-means clustering, the effectiveness of various approaches was evaluated by computing the Silhouette Score. Subsequently, an N-gram analysis was employed to identify the prevalent terms in the context, thus defining the distinctive contextual aspects associated with a specific domain. Finally, subject distributions of the papers were examined using HDP-BERT-based topic modelling analysis. Because the coherence score determines which model is better depending on whether the score is greater than zero, the coherence score of our suggested system is high in comparison to other current models. Furthermore, by doing more thorough research on certain themes, the study's findings might inspire in-depth inquiries.

7.0 DECLARATIONS

7.1 Ethics approval and consent to participate

Not applicable

7.2 Consent for publication

Not applicable

7.3 Availability of data and materials

Publically available dataset available in this link https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge.

7.4 Competing interests

The authors declare that they have no competing interests.

7.5 Funding

No funds, grants, or other support was received.

7.6 Authors' contributions

All authors equally contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

7.7 Acknowledgements
Not applicable

REFERENCES

[1] F. Gurcan, G. G. M. Dalveren and M. Derawi, "Covid-19 and E-Learning: An Exploratory Analysis of Research Topics and Interests in E-Learning During the Pandemic," in IEEE Access, vol. 10, pp. 123349-123357, 2022, doi: 10.1109/ACCESS.2022.3224034.
[2] H. Choi and Y. Ko, "Using Adversarial Learning and Biterm Topic Model for an Effective Fake News Video Detection System on Heterogeneous Topics and Short Texts," in IEEE Access, vol. 9, pp. 164846-164853, 2021, doi: 10.1109/ACCESS.2021.3122978.
[3] Zoya, S. Latif, F. Shafait and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," in IEEE Access, vol. 9, pp. 127531-127547, 2021, doi: 10.1109/ACCESS.2021.3112620.
[4] M. Bewong et al., "DATM: A Novel Data Agnostic Topic Modeling Technique with Improved Effectiveness for Both Short and Long Text," in IEEE Access, vol. 11, pp. 32826-32841, 2023, doi: 10.1109/ACCESS.2023.3262653.
[5] Y. Huang, R. Wang, B. Huang, B. Wei, S. L. Zheng and M. Chen, "Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model," in IEEE Access, vol. 9, pp. 108131-108143, 2021, doi: 10.1109/ACCESS.2021.3101565.
[6] C. -D. Curiac and M. V. Micea, "Identifying Hot Information Security Topics using LDA and Multivariate Mann-Kendall Test," in IEEE Access, vol. 11, pp. 18374-18384, 2023, doi: 10.1109/ACCESS.2023.3247588.
[7] X. Tan, M. Zhuang, X. Lu and T. Mao, "An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model," in IEEE Access, vol. 9, pp. 15860-15871, 2021, doi: 10.1109/ACCESS.2021.3052566.
[8] M. Duan, Q. Li and L. Xiao, "Topic-extended Emotional Conversation Generation Model Based on Joint Decoding," in IEEE Access, vol. 9, pp. 89934-89940, 2021, doi: 10.1109/ACCESS.2021.3090435.
[9] X. Sun and B. Ding, "Neural Network with Hierarchical Attention Mechanism for Contextual Topic Dialogue Generation," in IEEE Access, vol. 10, pp. 4628-4639, 2022, doi: 10.1109/ACCESS.2022.3140820.
[10] M. A. Razzaque, "Enabling Efficient and Scalable Service Search in IoT with Topic Modeling: An Evaluation," in IEEE Access, vol. 9, pp. 53452-53465, 2021, doi: 10.1109/ACCESS.2021.3071009.
[11] J. Rashid et al., "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," in IEEE Access, vol. 7, pp. 146070-146080, 2019, doi: 10.1109/ACCESS.2019.2944973.
[12] R. Devika, S. Vairavasundaram, C. S. J. Mahenthar, V. Varadarajan and K. Kotecha, "A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data," in IEEE Access, vol. 9, pp. 165252-165261, 2021, doi: 10.1109/ACCESS.2021.3133651.
[13] Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, Thoudam Doren Singh, "Prediction of research trends using LDA based topic modeling,"Global Transitions Proceedings, Volume 3, Issue 1,2022, Pages 298-304, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2022.03.015.
[14] A. Bannayeva and M. Aslanov, "Development of the N-gram Model for Azerbaijani Language," 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), Tashkent, Uzbekistan, 2020, pp. 1-5, doi: 10.1109/AICT50176.2020.9368645.
[15] J. Xue, X. Tang and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," in IEEE Access, vol. 9, pp. 125403-125415, 2021, doi: 10.1109/ACCESS.2021.3110683.
[16] Y. Zhang and Z. Rao, "n-BiLSTM: BiLSTM with n-gram Features for Text Classification," 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2020, pp. 1056-1059, doi: 10.1109/ITOEC49072.2020.9141692.
[17] X. She, J. Chen and G. Chen, "Joint Learning with BERT-GCN and Multi-Attention for Event Text Classification and Event Assignment," in IEEE Access, vol. 10, pp. 27031-27040, 2022, doi: 10.1109/ACCESS.2022.3156918.
[18] N. Khan, C. Agrawal and H. Yadav, "An Effective Compressive Sensing based N-gram Approach for plagiarism detection," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-7, doi: 10.1109/IDEA49133.2020.9170683.
[19] X. Lin, M. Liu and J. Zhang, "A Top-Down Binary Hierarchical Topic Model for Biomedical Literature," in IEEE Access, vol. 8, pp. 59870-59882, 2020, doi: 10.1109/ACCESS.2020.2983265.
[20] Zoya, S. Latif, F. Shafait and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," in IEEE Access, vol. 9, pp. 127531-127547, 2021, doi: 10.1109/ACCESS.2021.3112620.
[21] C. I. Eke, A. A. Norman and L. Shuib, "Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model," in IEEE Access, vol. 9, pp. 48501-48518, 2021, doi: 10.1109/ACCESS.2021.3068323.
[22] C. -O. Truică and E. -S. Apostol, "TLATR: Automatic Topic Labeling using Automatic (Domain-Specific) Term Recognition," in IEEE Access, vol. 9, pp. 76624-76641, 2021, doi: 10.1109/ACCESS.2021.3083000.
[23] P. Yang, Y. Yao and H. Zhou, "Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering," in IEEE Access, vol. 8, pp. 24734-24745, 2020, doi: 10.1109/ACCESS.2020.2969525.
[24] W. T. Alshammari and S. AlHumoud, "TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT with BiLSTM," in IEEE Access, vol. 10, pp. 91509-91523, 2022, doi: 10.1109/ACCESS.2022.3198955.
[25] X. Han and L. Wang, "A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information," in IEEE Access, vol. 8, pp. 96912-96919, 2020, doi: 10.1109/ACCESS.2020.2996642.

[26] C. Sharma, I. Batra, S. Sharma, A. Malik, A. S. M. S. Hosen and I. -H. Ra, “Predicting Trends and Research Patterns of Smart Cities: A Semi-Automatic Review using Latent Dirichlet Allocation (LDA),” in IEEE Access, vol. 10, pp. 121080-121095, 2022, doi: 10.1109/ACCESS.2022.3214310.

[27] X. Chen, P. Cong and S. Lv, “A Long-Text Classification Method of Chinese News Based on BERT and CNN,” in IEEE Access, vol. 10, pp. 34046-34057, 2022, doi: 10.1109/ACCESS.2022.3162614.