IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

IJARSCT

Impact Factor: 7.301

Volume 3, Issue 5, June 2023

# Exploring Latent Themes-Analysis of Various Topic Modelling Algorithms

**Reetesh Kumar Srivastava, Shalini Sharma, Dr. Piyush Pratap Singh**
School of Computer and Systems Sciences
Jawaharlal Nehru University, New Delhi, India

**Abstract***:* *This research explores the effectiveness of four common topic modelling methods for identifying latent themes and topics in unstructured text data: Latent Dirich- let Allocation(LDA), Non-Negative Matrix Factorization(NMF), Top2Vec, and BERTopic. Topic modelling is an essential method for gaining insights from mas- sive amounts of textual data. Top2Vec and BERTopic are recent approaches that use unsupervised neural networks to develop distributed representations of texts and words, whereas NMF and LDA are traditional techniques frequently utilised for topic modelling. This document gives a timeline of important advances in topic modelling, including the development of NMF and LDA, as well as many refinements and additions to LDA. According to the study's findings, BERTopic surpasses the other approaches, particularly in recognising overlapping and fine- grained subjects. This work emphasises the significance of text processing quality, the variety of subjects in the text, and the right selection of topic modelling methods in efficiently breaking down topics.*

**Keywords:** LDA, NMF, Top2Vec, BERTopic

## I. INTRODUCTION

"Natural language processing" (NLP) was established to make it possible for comput- ers to comprehend, interpret, and produce human language. Finding useful information and insights from vast amounts of unstructured text data is one of the main issues in NLP. Topic Modeling is an NLP technique that enables Researchers and practitioners to find latent themes and topics in a text corpus

LDA (Latent Dirichlet Allocation) and NMF (Non-Negative Matrix Factorization) are popular topic modelling methods. To depict the distributions of document topics and topic words, NMF divides a text data matrix into two lower-dimensional matrices. On the other hand, each document in LDA is a combination of latent themes, with each topic represented as a distribution of words in the corpus or text data[1][2].

Top2Vec is a relatively new topic modelling algorithm that uses an unsupervised neural network to learn distributed representations of texts and words. To map each document to a dense vector in a low-dimensional space, the technique first trains a document encoder. Then, based on the vector representations of the documents, to group related documents together a clustering technique is used. To determine the topic associated with each cluster, it extracts the most associated terms. It has been demonstrated that Top2Vec is extremely scalable and can efficiently handle massive volumes of text data.

BERT(Bidirectional Encoder Representations from Transformers), a pre-trained powerful language model, is the foundation of BERTopic, a more modern topic mod- elling approach. To extract contextualised embeddings for each document in the corpus, BERTopic first uses BERT, after that, a clustering method is used to com- bine related documents based on their embeddings[3]. To extract the topic associated with each cluster, it then determines which words are the most relevant. It has been shown that BERTopic performs better than conventional topic modelling techniques, especially when it comes to recognising overlapping and fine-grained topics.

This paper compares the effectiveness of four topic modelling algorithms(NMF, LDA, Top2Vec, and BERTopic) in identifying latent themes and topics within large sets of unstructured text data, providing researchers and practitioners with valuable insights.

DOI: 10.48175/IJARSCT-11635

## II. LITERATURE REVIEW

A topic is a group of dominant keywords that are representative that helps us to identify what the topic is all about. To segregate topics following factors are important-

- The level of text processing quality.
- The variety of topics(latent) present in the text.
- The proper choice of algorithm.
- Topic modelling, as an area of research that has grown significantly over time. Here is a quick timeline of some of the most significant advancements in this area.

### 2.1 NMF

Non-Negative Matrix Factorization is a matrix decomposition technique that has been widely used for topic modelling. The method was first introduced for topic modelling by Lee and Seung in 1999 in his paper ”Finding Structure in Time” [4]. However, its application to text data for topic modelling was by Lee and Seung, published in 2001[5].

Many different fields, including image processing, bioinformatics, speech recognition, and text analysis have used non-negative matrix factorization

### 2.2 LDA

LDA is a popular topic modelling algorithm that was introduced by David Blei et al. in a 2003 paper titled ”Latent Dirichlet Allocation[6]”

LDA was developed as a probabilistic generative model for text corpus/documents, It depicts each text document as a combination of topics, where each topic represents a distribution across words. Bayesian inference approach was used to estimate the model parameters, and the Gibbs sampling algorithm for posterior inference[6].

Due to its capacity to automatically identify latent themes or topics in huge text corpora, LDA has grown to become one of the most used topic modelling algorithms. It has been applied in various domains, such as document classification, social network analysis, sentiment analysis, recommender systems, search engines and bioinformatics. Over the years, many enhancements and advancements happened in the per- formance and scalability of LDA, as well as new extensions of the algorithm were developed. Some of them are -

### 2.3 Online LDA

A variant of LDA that enables the model to be updated incrementally as new doc- uments are made accessible. It is made for massive streaming datasets where it is impractical to compute the full model every time new data is received. Online LDA is computationally effective and scalable since it updates the model parameters using stochastic gradient descent[7].

### 2.4 Supervised LDA(sLDA)

In sLDA, a set of predetermined topics and link them to specific terms in the vocabu- lary, as opposed to regular LDA, which is an unsupervised method that infers themes only from the distribution of words in the texts. This prior information can be used in sLDA to enhance the subjects that are inferred. This model uses maximum likelihood procedure for parameter estimation, which uses parameter approximations in order to deal with difficult posterior expectations[8].

### 2.5 Dynamic Topic Models

In this model, it is assumed that the topics change gradually over time and that the topic distributions of related periods are comparable. So this method allows modelling dynamic nature and identifying how they change over a period of time [9].

DOI: 10.48175/IJARSCT-11635

### 2.6 Top2Vec

Top2Vec is a distributed topic vector model that determines topic vectors by using dense regions of document vectors. Topic vectors are derived as the centroids of dense areas of document vectors, where the total number of dense areas of documents in the semantic space is considered to match the total number of observable subjects. The model generates simultaneous documents, word vectors and embedded topics to repre- sent similarity. Top2vec and probabilistic models, like LDA, differ mostly in modelling topics. Since LDA views topics as word distributions, it is possible for uninformative words to have a high probability in topics as a result of their predominance in the corpus.

Top2vec does not require stop-word removal, lemmatization, stemming, or a priori knowledge of the number of topics in order to learn effective topic vectors. This model represents topics shared among documents and the nearest word to the topic vector that best describes the topic[10].

### 2.7 BERTopic

BERTopic is the latest topic modelling algorithm that is based upon pre-trained lan- guage model BERT, to generate topics from text data. It is built on the idea of embeddings to group related documents into topics. Each topic is represented by a set of keywords. [11].

BERTopic has been shown to outperform other traditional topic modelling algorithms LDA and NMF which were discussed above, on various data sets and evaluation metrics.

## III. DATA PREPROCESSING

Preprocessing is an essential phase in topic modelling since it aids in preparing raw text data for modelling by cleaning and transforming it. Following are some common topic modelling preprocessing steps

### 3.1 Data Transformation

- Data transformation is a process of converting raw textual data into a format suitable for analysis.
- Tokenization- is the process of breaking down the text into smaller parts, such as words or phrases, in order to analyse them independently.
- Stemming and Lemmatization- The process of reducing words to their basic form[12].
- Stopword removal: Removing frequent terms like ”the”, ”and”, ”is”, and so on because they don’t provide much value to subject modelling[13].

### 3.2 Feature Extraction

Identifying key details in the incoming text data and transforming it into a format appropriate for analysis is a critical step in topic modelling. The goal is to extract the text’s main qualities while reducing noise and extraneous information. Feature extraction techniques in topic modelling use ”Bag-of-Words” , ”term frequency-inverse document frequency (TF-IDF)” and word embedding.

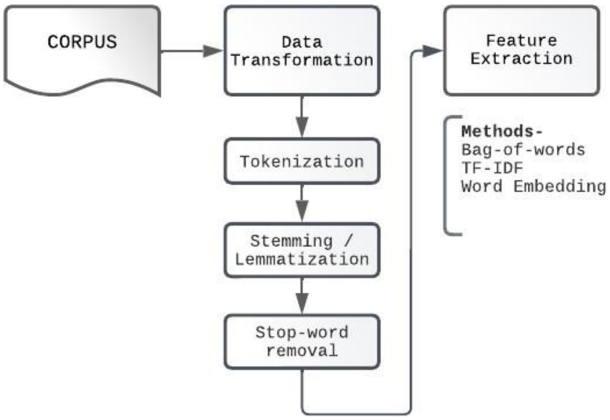

**Fig. 1** Flowchart representing step taken for prepossessing of data.

DOI: 10.48175/IJARSCT-11635

BoW is a simple and widely used feature extraction technique in NLP. It depicts a document as a bag (an unordered collection) of its words, that ignores syntax and word order and focuses solely on the frequency of each word. This produces a sparse matrix in which each column represents a word and each row represents a document, and the cells hold the frequency of the associated word in the related document[14].

TF-IDF is a BoW technique extension. It considers not just the frequency of each word, but also how essential a word is to the corpus as a whole. This is accomplished by multiplying the frequency of each word by its inverse document frequency (IDF), which measures how frequently the word appears in the corpus as a whole. The logarithm of the total number of documents divided by the number of documents that contain the word yields the IDF value. This gives less weight to terms that appear most frequently and gives more weight to terms that appear less frequently in documents, resulting in a more precise representation of the unique characteristics of each document [15][16].

## IV. COMPARATIVE ANALYSIS OF VARIOUS TOPIC MODELLING ALGORITHMS

LDA, NMF, Top2Vec, and BERTopic are popular topic modelling techniques that were discussed so far, and used in natural language processing (NLP). Here is a comparison of these techniques:

### 4.1 Approach

- LDA and NMF are probabilistic topic modelling techniques that model topics as a distribution of words[17].
- Top2Vec and BERTopic are embedding-based approaches that represent topics by dense vectors in semantic space

**Table 1** Comparison of Topic modelling algorithms.

| **Algorithm** | **LDA** | **NMF** | **Top2Vec** | **BERTopic** |
|---|---|---|---|---|
| **Approach** | Probabilistic | Probabilistic | Embedding | Embedding |
| **Input** | Bag-of-Word | Bag-of-Word | TF-IDF | cTF-IDF |
| **Preprocessing** | Required | Required | Not required | Not required |
| **Model Complexity** | Relatively Less | Relatively Less | High Complex | High Complex |
| **Year** | 2003 | 2000 | 2020 | 2022 |
| **Python Library** | sklearn/gensim | sklearn/gensim | Top2Vec | BERTopic |

This table shows comparison algorithms based on various parameters

### Input

- LDA and NMF require BoW(Bag-of-Words) representation of the text corpus.
- Top2Vec and BERTopic can work with raw text and preprocessed text. It uses TF-IDF & cTF-IDF respectively in most cases.

### Preprocessing

- LDA and NMF require preprocessing steps such as removing stop words, stemming, and lemmatization.
- Top2Vec and BERTopic can work with raw text or preprocessed text and do not require preprocessing steps.

### Model Complexity

- LDA and NMF are relatively simple models with fewer parameters as compared to others.
- Top2Vec and BERTopic are more complex models that require the training of large neural networks.

## V. CONCLUSION

To conclude, Natural Language Processing (NLP) approaches, particularly Topic Mod- elling, are crucial in helping researchers and practitioners extract valuable insights from enormous amounts of unstructured text

DOI: 10.48175/IJARSCT-11635

data. The conventional methods for topic modelling have been Latent Dirichlet Allocation and Non-Negative Matrix Factorization. Top2Vec and BERTopic, on the other hand, are more recent techniques that have shown greater performance, particularly in finding overlapping and fine-grained topics. The success of the topic modelling method is determined by aspects such as text processing quality, topic variety, and algorithm selection. Topic Modelling has significantly progressed, with numerous breakthroughs and extensions such as online LDA, supervised LDA, and Dynamic Topic Models

## REFERENCES

[1] Egger, R. & Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology 7 (2022). [Online; accessed 2023-05-06].

[2] Rodriguez-Garcia, P., Li, Y., Lopez-Lopez, D. & Juan, A. A. Strategic decision making in smart home ecosystems: A review on the use of artificial intelligence and internet of things. Internet of Things 100772 (2023).

[3] Lyu, Y. Dockerized knowledge-oriented multi-modal social event detection system, 1–6 (IEEE, 2022).

[4] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999). [Online; accessed 2023-05-09].

[5] Lee, D. & Seung, H. S. Leen, T., Dietterich, T. & Tresp, V. (eds) Algorithms for non-negative matrix factorization. (eds Leen, T., Dietterich, T. & Tresp, V.) Advances in Neural Information Processing Systems, Vol. 13 (MIT Press, 2000). URL https://t.ly/q 17.

[6] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).

[7] Hoffman, M. D., Blei, D. M. & Bach, F. Online learning for latent dirichlet allocation, NIPS'10, 856–864 (Curran Associates Inc., Red Hook, NY, USA, 2010).

[8] Blei, D. M. & McAuliffe, J. D. Supervised topic models, NIPS'07, 121–128 (Curran Associates Inc., Red Hook, NY, USA, 2007).

[9] Blei, D. M. & Lafferty, J. D. Dynamic topic models, ICML '06, 113–120 (Asso- ciation for Computing Machinery, New York, NY, USA, 2006). URL https://doi.org/10.1145/1143844.1143859.

[10] Angelov, D. Top2vec: Distributed representations of topics (2020). 2008.09470.

[11] Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure (2022). 2203.05794.

[12] Vijayarani, S., Ilamathi, M. J., Nithya, M. et al. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks 5, 7–16 (2015).

[13] Rahimi, Z. & Homayounpour, M. M. The impact of preprocessing on word embedding quality: A comparative study. Language Resources and Evaluation 57, 257–291 (2023).

[14] Qader, W., M. Ameen, M. & Ahmed, B. An overview of bag of words;importance, implementation, applications, and challenges, 200–204 (2019).

[15] Aizawa, A. An information-theoretic perspective of tf–idf measures. Information Processing Management 39, 45–65 (2003). URL https://www.sciencedirect.com/ science/article/pii/S0306457302000213.

[16] Gefen, D. et al. Identifying patterns in medical records through latent semantic analysis. Communications of the ACM 61, 72–77 (2018).

[17] Mardones-Segovia, C., Wheeler, J. M., Choi, H.-J., Wang, S. & Cohen, A. S. Model selection for latent dirichlet allocation in assessment data. Psychological Test and Assessment Modeling 65, 3–35 (2023).

DOI: 10.48175/IJARSCT-11635