Topics in Motion: A Review of Evolving Methods in Topic Modeling

Vijeta Kanwar,Sakshi,Rajvi Rana

Indira Gandhi Delhi Technical University for Women123

vijeta149btit24@igdtuw.ac.in,sakshi121btit24@igdtuw.ac.in,rajvi113btit24@igdtuw.ac.in

# Abstract

*Topic modeling has emerged as a powerful technique for uncovering thematic structures and identifying modern trends in a large text corpora. We explored key algorithms of topic modeling and looked into LDA and its extensions like OLDA. We discussed inference methods, like Gibbs Sampling and Variational Bayes, and their respective resources and requirements. Furthermore, we examined OLDA (Online LDA), which enables real-time interpretation. By analyzing and comparing existing research, we provide insights into strengths, limitations, and practical applications of these approaches. This review aims to guide future work by highlighting gaps in explainable AI, interpretability, scalability, and real-time data processing.*

**Keywords:** Topic Modelling,Linear discriminant analysis (LDA),People's Linguistic Survey of India (PLSI),Bagging,Bayes Model

## Introduction to Topic Modeling

* 1. **Overview of Topic Modeling**

Topic modeling is a powerful tool for identifying latent theme patterns in large volumes of documents. These methods reveal the text's hidden topics by identifying word clusters that frequently occur together. The basic tenet is that each text is a combination of several themes, each of which is determined by a specific word distribution. This ability extends beyond mere keyword analysis in that it takes into account the context of word occurrences so that semantic knowledge may be extracted and non-obvious relationships in the data identified. This renders topic modeling a priceless resource for many applications, such as information retrieval, text categorization, and content recommendation, where the underlying meaning of text is of the essence. Finally, topic modeling translates massive amounts of unstructured text into a cleaner and more interpretable form, making it easy to browse, search, and summarize document collections.

## Evolution of Topic Modeling Techniques Evolution of topic modelling

From mathematical curiosity to a powerful tool for revealing hidden topics in massive text corpora, topic modeling has undergone significant development. Latent Semantic Analysis (LSA) was first introduced by Deerwester et al. in the 1990s. They employed Singular Value Decomposition (SVD) to find patterns in the links between terms and documents. However, LSA lacked a probabilistic basis and was limited in its ability to handle uncertainty. This

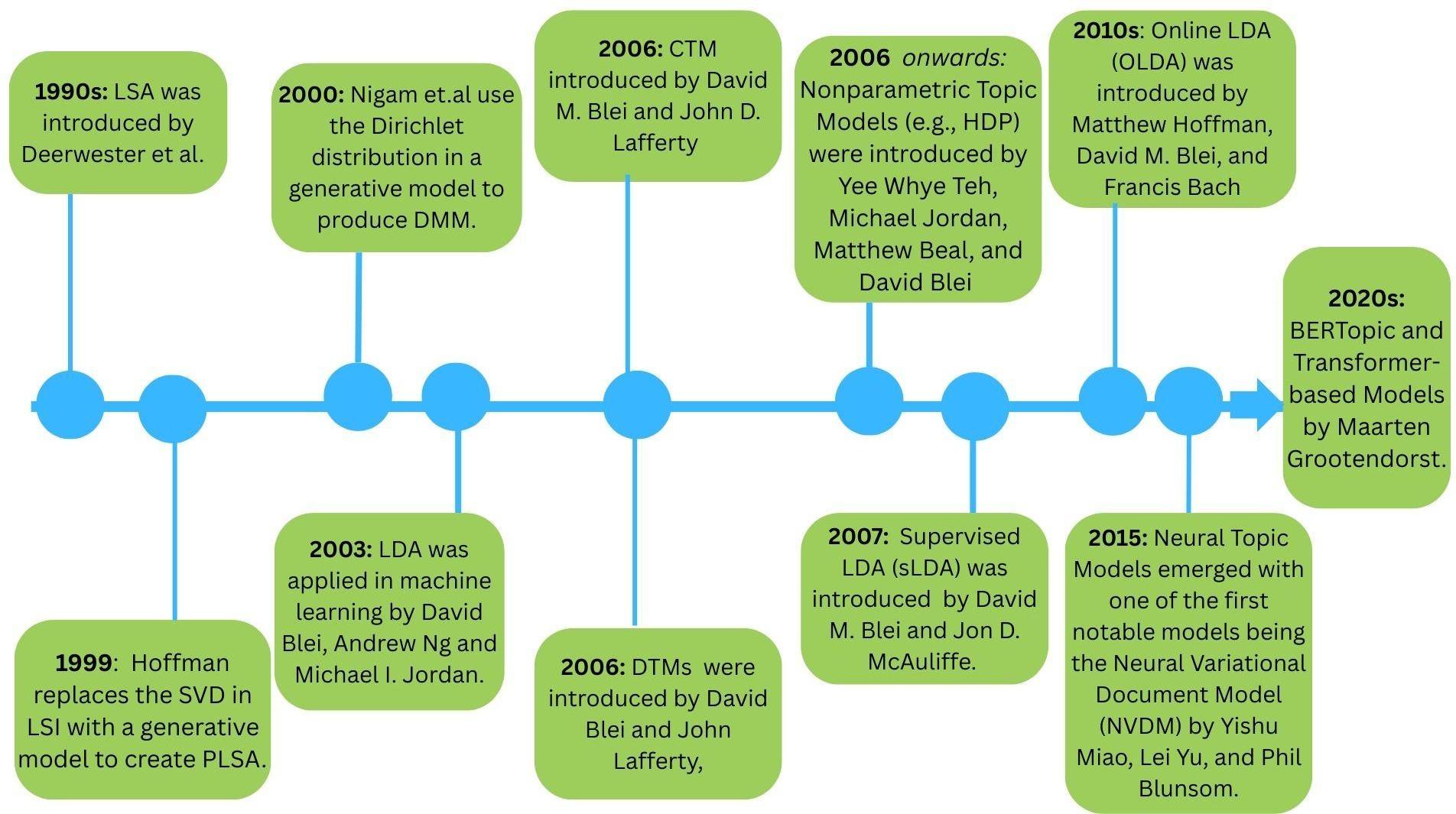
drawback was overcome in 1999 when Thomas Hofmann developed Probabilistic Latent Semantic Analysis (PLSA), a generative probabilistic model that offered a more sophisticated understanding of topic distribution in place of SVD.

Dirichlet-based models were introduced in the early 2000s, signaling a turning point in the journey of topic modeling. The Dirichlet Multinomial Mixture (DMM) model was developed by Nigam et al. in 2000, using the Dirichlet distribution. This opened the way for the Latent Dirichlet Allocation (LDA), which was introduced in 2003 by David Blei, Andrew Ng, and Michael I. Jordan. LDA quickly became popular due to its probabilistic foundation and ability to model each document as a collection of topics. LDA was further improved in the following years, with Correlated Topic Models (CTM) and Dynamic Topic Models (DTM) established in 2006 to simulate topic dependencies and topic change over time, respectively. This was followed in 2007 by supervised LDA, which made topic modeling useful for classification and regression problems.

With the growth in data amount and complexity, new approaches and tools were required. This led to the development of nonparametric topic models in 2006, such as the Hierarchical Dirichlet Process (HDP), which eliminated the need to specify the number of topics before receiving the data. To address large-scale data, Matthew Hoffman presented Online LDA (OLDA) in the 2010s, which makes topic modeling scalable to streaming and big data contexts. Because of its efficiency, OLDA has become widely used in sectors such as news analytics and real-time social media monitoring.

The introduction of deep learning caused a significant shift in subject modeling. In 2015, neural topic models were developed, including the neural Variational Document Model (NVDM), which increased topic coherence by utilizing variational autoencoders.

Finally, in the 2020s, transformer-based models emerged, like Maarten Grootendorst's BEERTopic. These models create high-quality topics by combining contextual embeddings from transformers such as BERT with clustering algorithms (for example, HDBSCAN, UMAP). BERTopic and similar models have emerged as the new standard for applications that require brief messages with nuanced semantics.



## Classical Approaches to Topic Modelling

* 1. **Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (pLSA)** Latent Semantic Analysis (LSA) is a technique used in natural language processing, specifically in information retrieval, to analyze relationships between documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent words and columns represent paragraphs) is constructed from a large piece of text, and a singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns. Words are then compared by taking the cosine of the angle between the two vectors (or the dot product between the normalized vectors). Values close to 1 represent very similar words, while values close to 0 represent dissimilar words.

Probabilistic Latent Semantic Analysis (pLSA) is a statistical technique for analyzing two-mode and co-occurrence data. In the information retrieval community, it has become

popular as a topic model alternative to latent semantic indexing, which overcomes some of LSI's shortcomings, most notably the lack of a sound probabilistic model. pLSA equates documents in the collection with mixtures of topics and each topic with a mixture of words. pLSA has its foundation in the multinomial distribution, while LSA is based on SVD.

## Latent Dirichlet Allocation (LDA)

LDA is a probabilistic approach that depicts subjects as word distributions and documents as topic mixtures [5]. Each document in a corpus is a composite of many topics, and each topic is defined by a probability distribution across all of the vocabulary's terms. This is how LDA operates. Given the observed word counts in the texts, the model uses a procedure called Bayesian inference to infer these topic distributions for each document and word distributions for each subject.

For natural language processing, it is an unsupervised machine learning method that finds connections and correlations in textual data [12]. LDA is appropriate for data analysis since it functions without any predetermined labels or groups.

LDA helps in the reduction of a vast collection of documents into a topical subspace, which facilitates comprehension and explanation [13]. LDA lowers the dimensionality of the data by depicting documents as collections of themes. This enables visitors to easily understand the general information and explore the text while capturing the key ideas in the documents.

When working with high-dimensional text data, where normal analytic techniques could be computationally costly or inefficient, this dimensionality reduction is especially helpful.

## Extension of LDA

* 1. **Dynamic topic model (DTM)**

The dynamic topic model (DTM) is an extension of LDA that captures the evolution of topics over time in a sequentially arranged corpus of documents, making it easy to visualize the topic trend. It works impressively for extracting topics from data that change slowly over

time. It replaces the Dirichlet distribution of topics with a sequence of Gaussian distributions, each representing the topic distributions for a different period.

DTM divides a dataset into periods, wherein only documents from a given period are exchangeable. The topics of one time period are conditioned on the topics from the previous time period, leading to topics that seemingly evolve through time. It can also predict future topics given its current topic distribution.

## Hierarchical Dirichlet Process (HDP)

The Hierarchical Dirichlet Process (HDP) is a nonparametric Bayesian model that allows the number of topics to be learned from the data rather than being predefined. HDP is a hierarchical Bayesian algorithm [**[2]**](https://answerthis.io/257914/answerthis/answer/d7e592392105a811#chunk_f8c8211398092f157c3219f83962742474faa1a4b520dd096bb7ee0d7476e6fa_25). It is a topic model that can learn the number of topics from the data [**[1]**](https://answerthis.io/257914/answerthis/answer/d7e592392105a811#chunk_0828e48a95718ec6a13eb6decbb4a0f7b6cb20368c1aef32b0c418d70f570dde_55).

Online HDP outperforms online LDA and batch HDP significantly on test data sets [**[1]**](https://answerthis.io/257914/answerthis/answer/d7e592392105a811#chunk_0828e48a95718ec6a13eb6decbb4a0f7b6cb20368c1aef32b0c418d70f570dde_55). Online algorithms are an important step in the evolution of topic models. However, they are still limited by the number of documents they can handle.

## Topics over Time (TOT)

Topics over Time (TOT) captures word co-occurrences and localization in continuous time. TOT, like cDTM, does not require time to be discretized, opting for a continuous distribution that is less computationally intensive than the original DTM.

## Continuous Dynamic Topic Model (cDTM)

The continuous version of the DTM, cDTM, does not require time to be separated. It uses a continuous distribution that is less computationally intensive than DTM. cDTM probabilistically selects a time frame for each topic before inferring the word distribution, allowing faster inference of a more finely distributed timeline of topics.

## Topic Tracking Model (TTM)

TTM was created for analyzing consumer purchase behavior on e-commerce websites. DTMs use a Dirichlet prior, which represents a probability distribution, but TTM, instead of simply passing a Dirichlet prior, modifies how that influence is applied, focusing on how items are grouped.

## Multi-Topic Trend Model (MTTM)

MTTM works hierarchically, starting with the smallest period and building up into bigger periods. This hierarchical structure allows users to “zoom in” on certain periods for a particular topic.

## Topic Flow Model (TFM)

The Topic Flow Model (TFM) reduces noise in topics in a temporal setting and tracks the evolution of topics. In addition to the emerging terms detected at each period, TFM seeks to confirm the continuing existence of previous topics.

## Online LDA

Online LDA is another form of LDA designed to handle text streams in a more dynamic and memory-efficient manner. Unlike traditional LDA, which operates in a way that requires the whole document, OLDA processes data incrementally, updating the topic as new documents arrive. It does this by using the previously learned topic-word distributions before the new incoming data, which helps the model to evolve without starting from the first document. This approach is particularly beneficial for real-time applications like news feeds, social media, or digital archives. OLDA can detect emergency topics and track how topics evolve, all while significantly reducing the computational and memory clustering of standard LDA. (rs4)

## Labeled LDA

Labeled LDA is a supervised extension of traditional LDA, which is designed for multi-labeled text or documents. Labeled LDA associates each topic directly with a

predefined label. This one-to-one correspondence ensures that every topic corresponds to a known category, making the resulting model more interpretable and suitable for document classification and tag-specific snippet extraction. In this model, only the labels assigned to a document are used to generate its words, efficiently constraining the generative process and improving topical relevance. This approach excels in scenarios where documents come with multiple known tags, such as social bookmarking platforms or annotated datasets. (Rs7)

## Contemporary Advances

It represents a shift from traditional probabilistic algorithms to more dynamic, semantically modified models that handle short, sparse, or noisy text data.

## BERTopic

It uses BERT-based contextual embeddings, dimensionality reduction (e.g., KernelPCA), and clustering (e.g., K-means) to generate logical topics, particularly excelling in social media and customer feedback scenarios. (rs12)

## Stochastic Block Models (SBM)

It is another advancement in the topic modeling, where documents and words are treated as nodes in a bipartite graph. This method reframes topic modeling as a network community detection problem, automatically inferring the number of topics and allowing for hierarchical structures. (rs14)

## Neural Topic Model

Neural topic models, such as the Embedded Topic Model (ETM) and Neural Variational Document Model (NVDM), integrate deep learning techniques to embed words and topics into a shared semantic space. These models improve topic clarity and are especially effective on short texts.

## Topic Modeling in Social Media

* 1. **Challenges of Topic Modeling in Social Media Data**

Due to the nature of the data of social media, topic modeling faces unique challenges. Data on social media is often short, does not have any repeated word patterns, and changes over time. Additionally, it also contains a significant amount of noise, requiring topic models to filter out irrelevant information effectively. Social media facilitates the publishing of information in real time, meaning that the subject of today was probably not that of yesterday and will likely not be that of tomorrow.

A topic model for social media should be capable of handling short and scattered data and text, tracking how topics change over time, and also removing irrelevant content to create clear topics. Since social media has many different types of posts, no one model works best for every situation.

## Adapting Topic Models for Short and Noisy Texts

Several approaches have been developed to adapt topic models for short and noisy texts, which are common in social media data. Some models are upgraded versions of older models, such as the Dirichlet Multinomial Mixture (DMM). Many of the best models are adaptations of LDA to account for these modern problems. Graph-based approaches have also been used in certain settings to reduce noise and find subsets of topics, but these models could not find the full topic set and cannot perform the same document classification as generative models.

Non-negative matrix factorization can be used as a topic model by breaking text into parts that represent topics. Also, using tools like word embeddings can make older topic models work better on different kinds of text without needing much extra computational power.

## Incorporating Word Embeddings

Word embeddings, such as Word2Vec, have been incorporated into topic models to improve their performance on social media data. Gibbs-sampled Pseudo document-based Dirichlet multinomial mixture (GPUDMM), a topic modeling algorithm specially designed for short texts, like tweets and social media posts. Unlike the traditional methods, which struggle with short texts as they lack context, GPUDMM treats a group of short texts as a

pseudo-document, combining them to form a bigger chunk of data so it can better detect patterns and topics. In this model, when a word is sampled, a set of semantically similar words is added back to the chosen topic.

## Applications

* 1. **Social Media Observation (grade 12)**

Nigerian bank customers' tweets were analyzed using BERTopic. This methodology successfully categorized Short, informal messages into logical subjects, exposing the main difficulties such as ATM malfunctions, customer support problems, and feedback from mobile banking. For real-time social media data, this demonstrates how embedding-based topic models perform better than conventional LDA.

## Exploration of Text via Networks (rs14)

The stochastic block model (SBM) was used in a network-based topic modeling framework. In this approach, words and documents are treated as nodes in a bipartite graph to identify

topic hierarchies. It demonstrated enhanced adaptability and automatic topic number recognition in tests conducted on both synthetic and historical corpora.

## Neural Topic Modeling for Sparse Data (rs15)

Embeddings are used by the Neural Variational Document Model (NVDM) and Embedded Topic Model (ETM) to improve coherence and perform well in short text domains.

## Streaming Topic Detection (rs4)

One adaptive model for analyzing real-time test streams is called Online LDA (OLDA). When a new document is received, this method changes the topic distributions; the entire dataset need not be kept. This is useful for trend analysis, news tracking, and following the evolution of web data.

## Biomedical Text Categorization (rs8)

A hybrid system that classified biomedical literature by combining ensemble classifiers and optimal LDA. The approach is perfect for medical literature mining and healthcare informatics since it increases accuracy in classifying research abstracts about illnesses, medication side effects, and diagnostics.

## Digital Journalism Analysis (rs11)

The New York Times archives were subjected to LDA to trace nuclear energy developments starting around 1945. This demonstrated how topic models may be used to measure media framing across time, which would be useful for public policy analysis, political communication research, and journalism studies.

## Mapping Academic Research Trends (rs13)

For thirteen years, statistical research publications were analyzed using topic modeling. LDA disclosed thematic clusters, changing research interests, and productivity by region.

Understanding scholarly advancement across journals and fields requires the use of this tool.

## Multi-Label Classification in Social Platforms (rs7)

Labeled LDA links particular words to recognized labels. When evaluated on social bookmarking sites, the algorithm enhanced credit attribution and engaged tag-specific content summary and classification.

## Applications

Topic modeling is applied in various fields such as software engineering, political science, and medical science [**[5]**](https://answerthis.io/257914/answerthis/answer/f3606fc77a24a02a#chunk_2bfebabd07abde86ae19fd30434157f2283a70db994bb3752cb9240b6c9c5458). In software engineering, it can be used to analyze code, bugs, and discussions to understand the evolution of software systems and identify areas of interest. In political science, it can be used to analyze speeches, news articles, and social media posts to understand public opinion and political discussions. In medical science, it can be used to analyze patient records and clinical trial data to identify trends and patterns in disease diagnosis and treatment.

In marketing, using advanced data analysis helps to uncover hidden connections and insights from vast amounts of information, like customer interactions or online activity, customer reviews, social media posts, and market research reports, to understand customer preferences, identify market segments, and track brand sentiment. [**[8]**](https://answerthis.io/257914/answerthis/answer/f3606fc77a24a02a#chunk_f8f61832ac6be4ed16131fafd837308092bbb434dcd3df61ec344c86744e3e65)By identifying the main topics discussed by customers, marketers can gain valuable insights into their needs and desires, allowing them to tailor their products and marketing campaigns more effectively.

Topic modeling is used in healthcare to identify important themes in rheumatology research

[10] and analyze trends in antimicrobial resistance research [9]. Researchers can find new trends in disease diagnosis, treatment, and prevention by using topic modeling to analyze the large body of medical literature. Additionally, it can assist in identifying important themes in particular fields of medical research, like rheumatology or antibiotic resistance, enabling researchers to concentrate their efforts on the most promising areas of study.

Social media can be a valuable tool for understanding how the public feels about food security issues. By analyzing posts on social media platforms, researchers can gain insights into public concerns, opinions, and perceptions related to food security. This information can inform policy decisions and help create more effective communication strategies.

**Table of Key Findings from Reviewed Papers**

The table below explores how topic modeling techniques have evolved over time, highlighting insights from a range of influential research papers. It reflects the journey of different models, methods, and ideas that have shaped the field across various applications.

| **Reference** | **Dataset Used** | **Algorithm/Model** | **Methodology** | **Result** | **Advantages** | **Limitation** | **Future Work** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Onan et al. (2016)** | Review- Polarity, Multi- Domain Sentiment,Irish-  Sentiment,Reviews  Dataset | NB, SVM, LR, RBF, KNN | Bootstrap aggregating  ,text mining, topic modeling | LDA  reduce dimensionality | Better interpretability, reduced feature space | Limited performance gains, fixed topic count, high complexity | Improve LDA, try supervised topic models, integrate deep learning |
| **Alsumait et al. (n.d.)** | Reuters-  21578,  NIPS  Papers | Online  LDA  (OLDA) | Incremental  updates, Gibbs  sampling,  KL- divergence  topic detection | Comparable to batch LDA,  real-time trend detection | Memory-  efficient,  tracks  topic  evolution | Sensitive  to weak  signals,  topic count  tuning  needed | Weighted KL-  metrics, prior  knowledge  usage,real-time  scalability |
| **Buenano- Fernandez et al. (2020)** | Teacher  survey  responses | LDA, text  network  modeling,  TF-IDF | Preprocessing, topic modeling,  network  construction | 12 topics  on student  retention  identified | Automates survey  analysis,  optimized  topic  extraction | Small  dataset,  manual  labeling,  no  demographics | Larger  datasets,  enhanced  preprocessing,  demographic  analysis |
| **Onan (2018)** | Oh5,  Oh10,  Oh15,  Ohscal,  Ohsume  d-400 | Swarm-  Optimized  LDA,  DEP | Preprocessing,  metaheuristic  for LDA,  ensemble  pruning | Higher  classification  accuracy, top  with BA-LDA,  Firefly-DEP | Optimized  biomedical topic  modeling,  better  accuracy | Computationally  heavy,  parameter  tuning  needed | Combine with  deep learning,  expand to  clinical apps,  improve  interpretability |
| **Rashid et al. (2019)** | Ohsumed,  GENIA,  Biotext,  Health  Tweets,  WSJ  Corpus | Hybrid  IDF,  Fuzzy K-  Means,  RPCA,  Discriminant  Classifier | Text  normalization, RPCA,  topic clustering | Higher  accuracy than  LDA/LSA,  good log-  likelihood | Redundancy  reduction,  accurate  topic  discovery | High  computational cost,  parameter  tuning  needed | Support  multilingual  data, boost  speed, use  deep learning |
| **Ramage**  **et al.**  **(2009)** | Delicious,  Yahoo Directory | Labeled  LDA (L-  LDA) | One-to-one label-topic  mapping,  Gibbs  sampling | Better  snippet  extraction than SVM,  competitive  distribution | Clear topic-  label links,  handles  multilabel data | Computational cost,  ignores  label  correlations | Add  correlated  models, apply  semi-  supervised  training |
| **Jelodar**  **et al.**  **(2017)** | CoPhIR,  Twitter, CiteSeer, Parliament Speeches | LDA,BTM,Corr-LDA,others | LDA survey, domain applications  , model comparisons | LDA  adopted widely, variants better in specific domains | Versatile, improves retrieval, extensible models | Expensive, lacks context, hyperparameter sensitivity | Deep learning fusion, scalable LDA, multilingual focus |
| **Blei et al. (2003)** | TRECAP, C.  Elegans, Reuters- 21578,  EachMo vie | LDA | Generative topic modeling, inference, perplexity tests | Outperforms pLSI  and mixture models | Multiple- topic handling, interpretability, generaliza  tion | Costly on large data,fixed topic number, BOW  limitations | Scalable inference, hierarchical topic support, correlation modeling |
| **Katz (n.d.)** | 100M-  word technical corpus | Poisson distribution mixtures | Burstiness modeling, probabilistic analysis | Burstiness affects word occurrence more than  frequency | Better prediction of repetitions  ,generalizable | Lacks semantic depth, tested only on technical data | Apply to broader genres, integrate semantics, enhance speech recognition |
| **Churchill & Singh (2022)** | Newspapers, Scientific Papers, Social Media | LDA +  extensions  Graph models,NMF,  Neural Topic models | Topic evolution review, model comparison s,  evaluation  matrices | Classic models  = good for structured data;  neural better for short-  text | Multi- domain insights, strong coverage of model  progress | High complexity  ,evaluation inconsistencies | Build hybrid models, real- time short- text modeling, better evaluations |

# Results and discussion

Topic modeling has evolved over the years from basic techniques like LSA to more advanced methods including neural networks and dynamic models .LDA is one of the most popular and versatile technique in topic modelling, with several extensions and variations.LDA's flexibility, accessibility and interpretability have made it a popular choice for various applications for the learnings.

Topic Modeling is a popular method that discovers the hidden theme and structure in unorganized biomedical text documents.These documents structure is used for searching, indexing and summarizing of documents. In machine learning, fuzzy techniques are widely used for biomedical image processing and text processing. (rs6)

The experimental results of topic modelling in biomedical classifications indicate that the proposed multiple classifier system(Bat Algorithm(BA), Diversity-Based Ensemble Pruning(DEP)) gives more accuracy than the conventional classification algorithms, ensemble learning,and ensemble pruning methods.It also reduces dimensionality of biomedical text data efficiently.

We have examined the predictive performance of classification algorithms (Naïve Bayes, support vector machines, logistic regression, radial basis function network and K-nearest neighbor algorithms) and ensemble learning methods (Bagging, AdaBoost, Random Subspace, voting and stacking)for text sentiment classification when LDA-based representation is utilized. (rs 5)

LDA can be used to filter out categories of texts that are not relevant from an overall sample.(rs11)

One of the applications of topic modelling using LDA (Latent Dirichlet Allocation) in journalism research, highlighting its value for analyzing large datasets. LDA helps the researchers easily identify the broad patterns and topic trends over time or across media. The interpretation process still requires effort but it's a cost-effective tool for initial analysis before more intensive content analysis.

Insights from topic modeling about either the formulation of suitable priors or the approximation of posterior distributions might catalyze the development of improved statistical methods to detect communities in networks. Furthermore, the traditional application of topic models in the analysis of texts leads to classes of networks usually not considered by community detection algorithms.(rs14)

LDA Models can be an efficient method for discovering latent group structure in large networks. The authors proposed a scalable Bayesian alternative based on LDA and graph to group discovery in a big real-world graph.Topic modelling in mage processing, Image classification and annotation derive an approximate inference and obtain algorithms based on variational ways as well as impressive approximations for annotating and classifying new images and extended supervised topic modeling (sLDA) to classification problems.

# Future work

Topic modeling will play an increasingly important role in helping us understand and navigate the ever-growing amount of text data . As the amount of text data continues to grow, topic modeling will become increasingly important for extracting meaningful insights and understanding complex phenomena. Interdisciplinary collaboration and ethical considerations will be playing crucial role for realizing the full potential of topic modeling . Interdisciplinary collaboration and ethical considerations are also important aspects for ensuring that topic modeling is used responsibly and ethically.In the future, bibliometric measures could be used to enhance topic descriptions. Additionally, analyzing the temporal evolution of topics could provide valuable insights into how themes develop and change over time.

As for the guidelines to take into account in future research,it would be desirable to work with a corpus with more data.Therefore it is suggested that there is a need to enrich the pre-processing phase with the application of the following techniques: disambiguation and part of speech tagging, entity

extraction (recognize proper names such as locations, organizations, or names of people) and n-gram detection (identify words that are grouped into a single term)( rs 9)

Semi-supervised models may be an important future direction that balances the computational cost and the cost of labeling training data. Topic models are also likely to become useful for generating features for different machine learning and NLP tasks. We are already beginning to see this for learning of complex dynamics like forced migration .Other possible applications include understanding conversation dynamics to detect malicious groups of users in social media networks, identification of different types of misinformation, and comparisons of different types of document collections (newspaper and social media as an example). These are all promising directions that

can further increase the impact of topic models.We can combine Topic models with matrix factorization methods to image understanding, tag assignment and semantic discovery from social image datasets (rs1)