(Ogunleye et al., 2023)**Topics in Motion: A Review of Evolving Methods in Topic Modeling**

Vijeta Kanwar, Sakshi, Rajvi Rana

Indira Gandhi Delhi Technical University for Women[123]
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 corpus. 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. With the growth in data amount and complexity, new approaches and tools were required. For this purpose, LDA parameters are optimized on abstracts and full-texts of articles published in two different scientific journals and the results obtained are compared with each other. This led to the development of nonparametric topic models in 2006 .By analyzing and comparing existing research, we provide insights into strengths, limitations, and practical applications of these approaches. In this paper we get motivate to provide a comprehensive review on topic modeling which includes a brief classification hierarchy as well as explanation of each method with algorithms like latent semantic analysis, probabilistic latent semantic analysis, non-negative matrix factorization and then finally the latent Dirichlet Allocation(LDA) is presented in great detail. 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), Dimensionality, Dynamic topic model (DTM)

Introduction to Topic Modeling

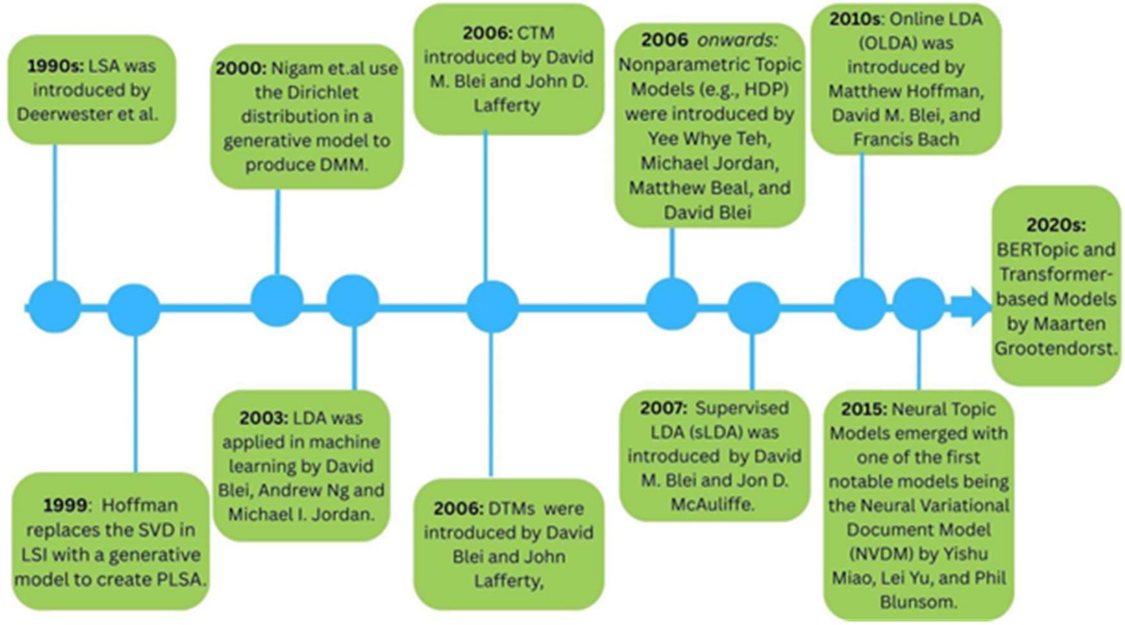
1. Overview of Topic Modeling

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.

2. **Evolution of Topic Modeling Techniques Evolution of topic modelling** (Churchill & Singh, 2022)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.



1990s: LSA was introduced by Deerwester et al.

1999: Hoffman replaces the SVD in LSI with a generative model to create PLSA.

2000: Nigam et.al use the Dirichlet distribution in a generative model to produce DMM.

2003: LDA was applied in machine learning by David Blei, Andrew Ng and Michael I. Jordan.

2006: CTM introduced by David M. Blei and John D. Lafferty

2006: DTMs were introduced by David Blei and John Lafferty,

2006 *onwards:* Nonparametric Topic Models (e.g., HDP) were introduced by Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei

2007: Supervised LDA (sLDA) was introduced by David M. Blei and Jon D. McAuliffe.

2010s: Online LDA (OLDA) was introduced by Matthew Hoffman, David M. Blei, and Francis Bach

2015: Neural Topic Models emerged with one of the first notable models being the Neural Variational Document Model (NVDM) by Yishu Miao, Lei Yu, and Phil Blunsom.

2020s: BERTopic and Transformer-based Models by Maarten Grootendorst.

Classical Approaches to Topic Modeling

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.

Two-mode and co-occurrence data can be analyzed statistically using Probabilistic Latent Semantic Analysis (pLSA). It has gained popularity in the information retrieval field as a topic model substitute for latent semantic indexing, addressing some of the drawbacks of LSI, chief among them being the absence of a reliable probabilistic model. pLSA assigns a variety of terms to each topic and a mixture of themes to the collection's documents. LSA is based on SVD, whereas pLSA is based on multinomial distribution.

2. **Latent Dirichlet Allocation (LDA)** (Jelodar et al., 2017)

LDA is a probabilistic approach that depicts subjects as word distributions and documents 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 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.

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 are conditioned on the topics from the previous time, leading to topics that seemingly evolve through time. It can also predict future topics given its current topic distribution.

2. **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]. It is a topic model that can learn the number of topics from the data [1].

Online HDP outperforms online LDA and batch HDP significantly on test data sets [1]. 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.

3. **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.

4. **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.

5. **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.

6. **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.

7. **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.

8. **Online LDA**

Another type of LDA that is intended to manage text streams more dynamically and memory-efficiently is online LDA. OLDA handles data gradually, updating the topic as new documents enter, in contrast to classical LDA, which requires the entire text. This allows the model to develop without having to start from the first document by applying the previously learnt topic-word distributions prior to the new incoming data. For real-time applications like news feeds, social media, or digital archives, this method is very advantageous. While drastically lowering the computational and memory clustering of conventional LDA, OLDA is able to identify emergency topics and monitor the evolution of subjects. [4]

9. **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. 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. [7]

Contemporary Advances

1. **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. [12]

2. **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. [14]

3. **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

5.1 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.

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.

1. **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.

5. 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.

2. **Neural Topic Modeling for Sparse Data** [4]

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.

3. **Streaming Topic Detection** [1]

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.

4. **Biomedical Text Categorization** [5]

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.

5. **Digital Journalism Analysis [11]**

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.

6. **Mapping Academic Research Trends [13]**

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.

7. **Multi-Label Classification in Social Platforms** [2]

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.

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.

eference	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.)	euters-1578, IPS apers	nline DA OLDA)	ncremental pdates, Gibbs ampling, L- divergence opic detection	omparable o batch DA, eal-time rend etection	Memory-fficient, racks opic volution	ensitive o weak ignals, opic count uning eeded	Weighted KL-metrics, prior nowledge sage,real-time calability
Buenano-Fernandez et al. (2020)	eacher urvey esponses	DA, text etwork modeling, F-IDF	reprocessing, opic modeling, etwork onstruction	2 topics n student etention dentified	utomates urvey nalysis, ptimized opic xtraction	mall ataset, manual abeling, o emographics	arger atasets, nhanced reprocessing, emographic nalysis
Onan (2018)	h5, h10, h15, hscal, hsume -400	warm-ptimized DA, EP	reprocessing, metaheuristic or LDA, nsemble runing	igher lassification ccuracy, top ith A-LDA, irefly-DEP	ptimized iomedical opic modeling, etter ccuracy	omputationaly eavy, arameter uning eeded	ombine with eep learning, xpand to linical apps, mprove nterpretability

Rashid et al. (2019)	hsumed, ENIA, iotext, ealth weets, WSJ orpus	ybrid DF, uzzy K-Means, PCA, iscriminant lassifier	ext ormalization, PCA, opic clustering	igher ccuracy han DA/LSA, ood log-ikelihood	edundancy eduction, ccurate opic iscovery	igh omputational ost, arameter uning eeded	upport multilingual ata, boost peed, use eep learning
amage t al. 2009)	elicious, ahoo irectory	abeled DA (L-DA)	ne-to-one abel-topic mapping, ibbs ampling	etter nippet xtraction han SVM, ompetitive istribution	lear topic-abel links, andles multilabel ata	omputationa cost, gnores abel orrelations	dd orrelated models, apply emi-upervised raining
elodar t 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, EachMovie	LDA	Generative topic modeling, inference, perplexity tests	Outperforms pLSI and mixture models	Multiple-topic handling, interpretability, generalization	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 techniques 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. [6]

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. [7]

LDA can be used to filter out categories of texts that are not relevant from an overall sample.[11]

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.[3]

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.

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10).

https://doi.org/10.1145/3507900

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2017). *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*. http://arxiv.org/abs/1711.04305

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences (Switzerland)*, *13*(2). https://doi.org/10.3390/app13020797

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2017). *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*. http://arxiv.org/abs/1711.04305

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences (Switzerland)*, *13*(2). https://doi.org/10.3390/app13020797