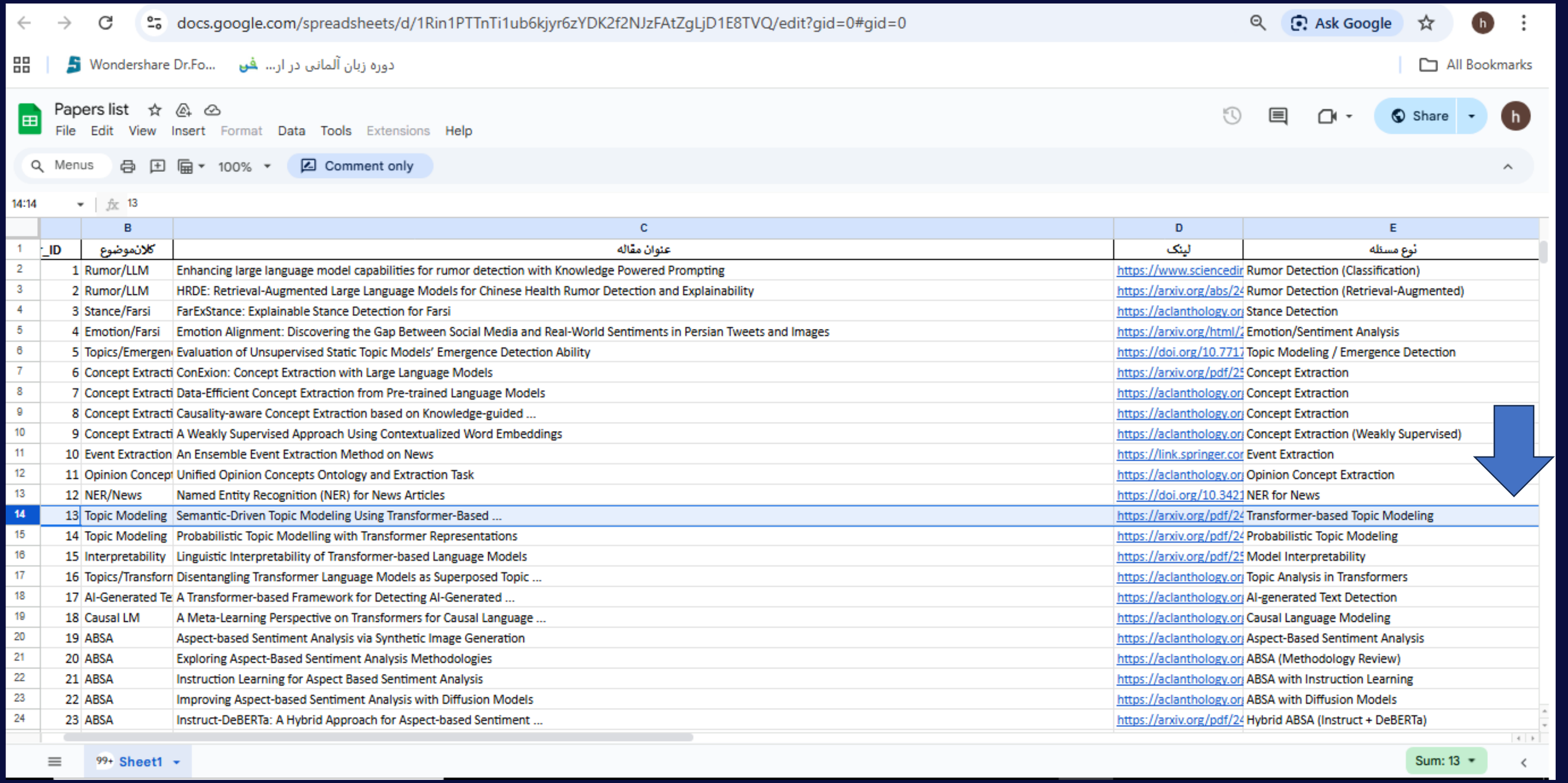
Transformer-based Topic Modeling

Natural Language Processing

Student: Hengameh Shah Ali
40307604

Transformer-based Topic Modeling



docs.google.com/spreadsheets/d/1Rin1PTTnTi1ub6kjyr6zYDK2f2NJzFAtZgLjD1E8TVQ/edit?gid=0#gid=0

Papers list

File Edit View Insert Format Data Tools Extensions Help

Share

Menus 100% Comment only

14:14 13

	B	C	D	E
1 _ID	کلان‌موضوع	عنوان مقاله	لینک	نوع مسئله
2	1 Rumor/LLM	Enhancing large language model capabilities for rumor detection with Knowledge Powered Prompting	https://www.sciencedir	Rumor Detection (Classification)
3	2 Rumor/LLM	HRDE: Retrieval-Augmented Large Language Models for Chinese Health Rumor Detection and Explainability	https://arxiv.org/abs/24	Rumor Detection (Retrieval-Augmented)
4	3 Stance/Farsi	FarExStance: Explainable Stance Detection for Farsi	https://aclanthology.or	Stance Detection
5	4 Emotion/Farsi	Emotion Alignment: Discovering the Gap Between Social Media and Real-World Sentiments in Persian Tweets and Images	https://arxiv.org/html/2	Emotion/Sentiment Analysis
6	5 Topics/Emergen	Evaluation of Unsupervised Static Topic Models' Emergence Detection Ability	https://doi.org/10.7717	Topic Modeling / Emergence Detection
7	6 Concept Extracti	ConExion: Concept Extraction with Large Language Models	https://arxiv.org/pdf/25	Concept Extraction
8	7 Concept Extracti	Data-Efficient Concept Extraction from Pre-trained Language Models	https://aclanthology.or	Concept Extraction
9	8 Concept Extracti	Causality-aware Concept Extraction based on Knowledge-guided ...	https://aclanthology.or	Concept Extraction
10	9 Concept Extracti	A Weakly Supervised Approach Using Contextualized Word Embeddings	https://aclanthology.or	Concept Extraction (Weakly Supervised)
11	10 Event Extraction	An Ensemble Event Extraction Method on News	https://link.springer.cor	Event Extraction
12	11 Opinion Concep	Unified Opinion Concepts Ontology and Extraction Task	https://aclanthology.or	Opinion Concept Extraction
13	12 NER/News	Named Entity Recognition (NER) for News Articles	https://doi.org/10.3421	NER for News
14	13 Topic Modeling	Semantic-Driven Topic Modeling Using Transformer-Based ...	https://arxiv.org/pdf/24	Transformer-based Topic Modeling
15	14 Topic Modeling	Probabilistic Topic Modelling with Transformer Representations	https://arxiv.org/pdf/24	Probabilistic Topic Modeling
16	15 Interpretability	Linguistic Interpretability of Transformer-based Language Models	https://arxiv.org/pdf/25	Model Interpretability
17	16 Topics/Transforn	Disentangling Transformer Language Models as Superposed Topic ...	https://aclanthology.or	Topic Analysis in Transformers
18	17 AI-Generated Te	A Transformer-based Framework for Detecting AI-Generated ...	https://aclanthology.or	AI-generated Text Detection
19	18 Causal LM	A Meta-Learning Perspective on Transformers for Causal Language ...	https://aclanthology.or	Causal Language Modeling
20	19 ABSA	Aspect-based Sentiment Analysis via Synthetic Image Generation	https://aclanthology.or	Aspect-Based Sentiment Analysis
21	20 ABSA	Exploring Aspect-Based Sentiment Analysis Methodologies	https://aclanthology.or	ABSA (Methodology Review)
22	21 ABSA	Instruction Learning for Aspect Based Sentiment Analysis	https://aclanthology.or	ABSA with Instruction Learning
23	22 ABSA	Improving Aspect-based Sentiment Analysis with Diffusion Models	https://aclanthology.or	ABSA with Diffusion Models
24	23 ABSA	Instruct-DeBERTa: A Hybrid Approach for Aspect-based Sentiment ...	https://arxiv.org/pdf/24	Hybrid ABSA (Instruct + DeBERTa)

Sheet1

Sum: 13

Problem Definition

Topic Modeling is one of the key challenges in Natural Language Processing (NLP), aiming to automatically discover hidden topics within a large collection of unlabeled textual documents. This emerging approach leverages transformers to provide a deeper understanding of textual content. In this model, each document is considered a mixture of multiple topics, and each topic is represented as a distribution over words/concepts. This approach helps uncover the latent semantic structure within vast amounts of textual data.

Transformer-based Topic Modeling is a novel method based on **transformer architectures** for discovering **hidden topical patterns** in large text collections. Thus, the core problem is extracting the thematic structure from texts without requiring labeled data, while achieving deep semantic understanding and high interpretability.

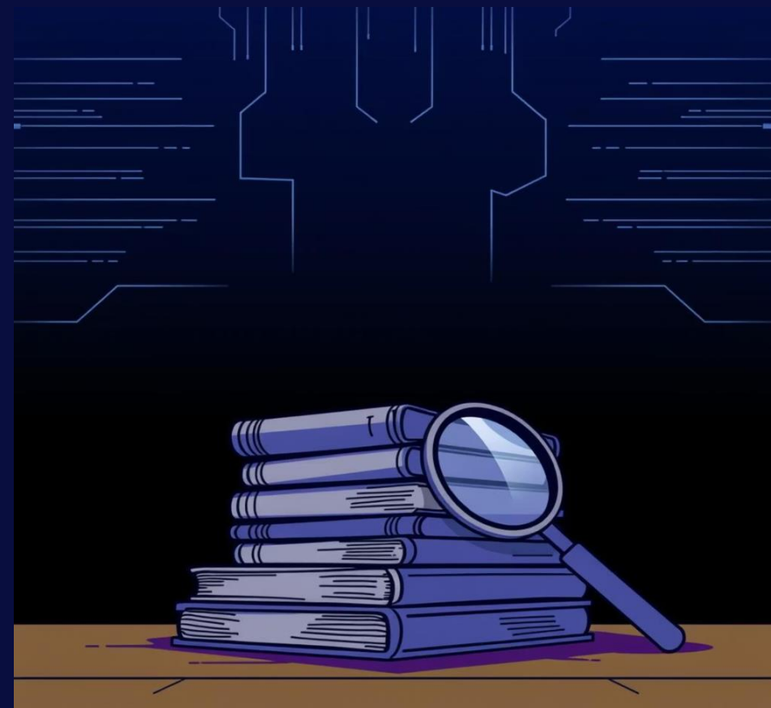
How can the latent thematic structures within a large collection of textual documents be **automatically discovered, organized, and interpreted**, while simultaneously:

Eliminating the need for **manual labeling**,

Preserving deep, **context-dependent meaning**,

Accounting for **complex relationships** between words and concepts, and

Ensuring **scalability** for massive datasets?



The Historical Evolution of NLP

❑ **The Statistical Era (2000–2010)**

LDA, pLSA, NMF

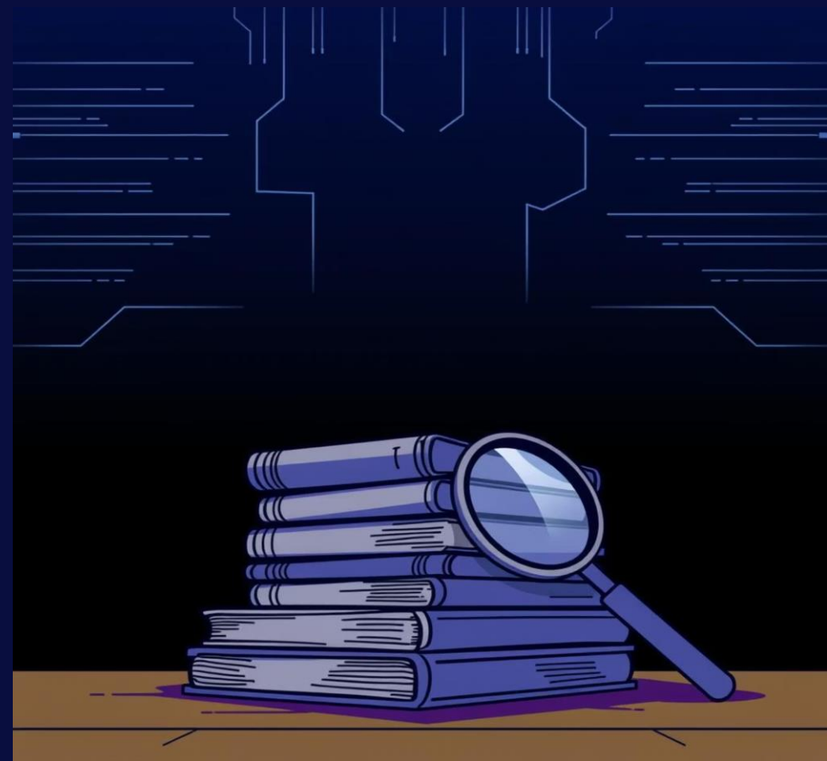
❑ **The Embedding-Based Era (2010–2018)**

Word2Vec + Clustering

❑ **The Transformer Revolution (2018–Present)**

Transformer-based Topic Modeling

Combining BERT's deep contextual understanding with advanced clustering techniques.



Comparison of Transformer-based Methods with Classical Methods

Roots: From Classical Models to the Need for Innovation

- In Transformer-based Topic Modeling, powerful language models such as BERT are used to extract deep semantic representations (Semantic Embeddings). This approach transforms the problem from analyzing word co-occurrence to clustering concepts in semantic vector space.

Classical Methods (e.g., LDA, PLSA, NMF)	Transformer-based Methods
Bag-of-Words-based modeling	Based on deep textual semantics while considering word context
Treating words independently	Modeling inter-word relationships (Self-Attention)
Requires a predetermined number of topics	Automatic topic extraction
Poor performance on short, multilingual, and specialized texts due to a lack of deep semantic understanding	Effective for both long and short texts

Significance and applications

Significance:

- **Enhancement of topic extraction quality**
- **Ability to distinguish between different usages of a word**
- **Reduction in the need for human intervention**
- Elimination of the need for predefined keyword lists
- **Scalability for massive datasets**
- Capability to process millions of documents

Application Example	Context
Identifying market trends from news and social media	Marketing and Market Research
Tracking significant topics in the news	Media monitoring
Automated categorization of support tickets	Customer Support
Automated categorization of documents, emails, and reports	Organizational knowledge management
Discovering emerging research fields from scientific articles	Scientometrics

Importance and key transformations introduced

✓ **Transformer-based Topic Modeling** has introduced profound transformations in how texts are understood and analyzed

Multilingual

Native capability for effectively localizing and analyzing multilingual texts without requiring separate models

Short texts

Significantly stronger performance in analyzing short texts, which was a major weakness of classical models

Understanding meaning

Transition from word-level to concept-level understanding, with greater accuracy and depth

Scalability

Capability to scale for processing organizational data and vast volumes of text

Interpretability

Providing results with higher interpretability, which is critical for decision-making

Real-world examples

✓ This technology is utilized across various industries and helps organizations gain valuable insights from their textual data

Media and social network analysis

Discovery of trends, narratives, and media crises; monitoring of public opinion. Example: Bloomberg and Reuters

Telco and Tech-Co

Analysis of customer complaints, support tickets, and call center conversations; discovery of recurring network issues and hidden grievances. Example: Vodafone and Telefónica

Organizational document analysis

Automatic categorization of reports, resolutions, and administrative correspondence; discovery of key policy-making topics

Customer feedback and market analysis

Review Mining and Voice of the Customer (VoC). Example: Amazon Review Analysis and Airbnb Feedback Mining

Health, legal, and finance

Analysis of medical articles, judicial opinions, and financial reports. Example: PubMed Topic Discovery and Legal Case Clustering

In classical models, **"bank loan," "credit facilities,"** and **"financial lottery"** were treated as separate topics

With transformers, these concepts are identified as a single topic: **financial services**, due to their semantic similarity

Practical Applications of Transformer-based Topic Modeling

❑ Advanced and intelligent content analysis systems

Advanced social media monitoring and analysis platforms such as Brandwatch and NetBase Quid, aimed at achieving a more nuanced and informed understanding of conversations.

❑ Intelligent customer support and complaint analysis systems

These systems can extract complex topics and subtopics from textual conversations (chat, email).

❑ Semantic search engines and organizational knowledge management systems

Next-generation research systems and digital libraries

Scientific databases such as Semantic Scholar and arXiv, designed for recommending articles with related concepts.

❑ Cybersecurity monitoring platforms

Cyber threat analysis systems: Used for processing large volumes of security reports, incident logs, and extracting **emerging threat topics**, **attack techniques (TTPs)**, and **indicators of compromise (IOCs)** from them.

Analysis of the relationship with key domain projects

- **Social networks and content analysis**

  Deep trend analysis: Detecting encoded speech on social networks. For example, identifying conversations centered around a specific domain.

  Online community identification: By analyzing topics in specific groups and channels, communities that have formed around particular subjects can be identified.

- **Cybersecurity and national security**

  Cybersecurity monitoring: Automated analysis of vast amounts of textual data from various sources (websites, forums, social networks) to uncover **blind spots of threats**.

  Analysis of communications for crime prevention: In the realm of national security, this method can be employed to analyze suspicious communications and extract **hidden plans and schemes**, while fully adhering to ethical considerations and privacy constraints.

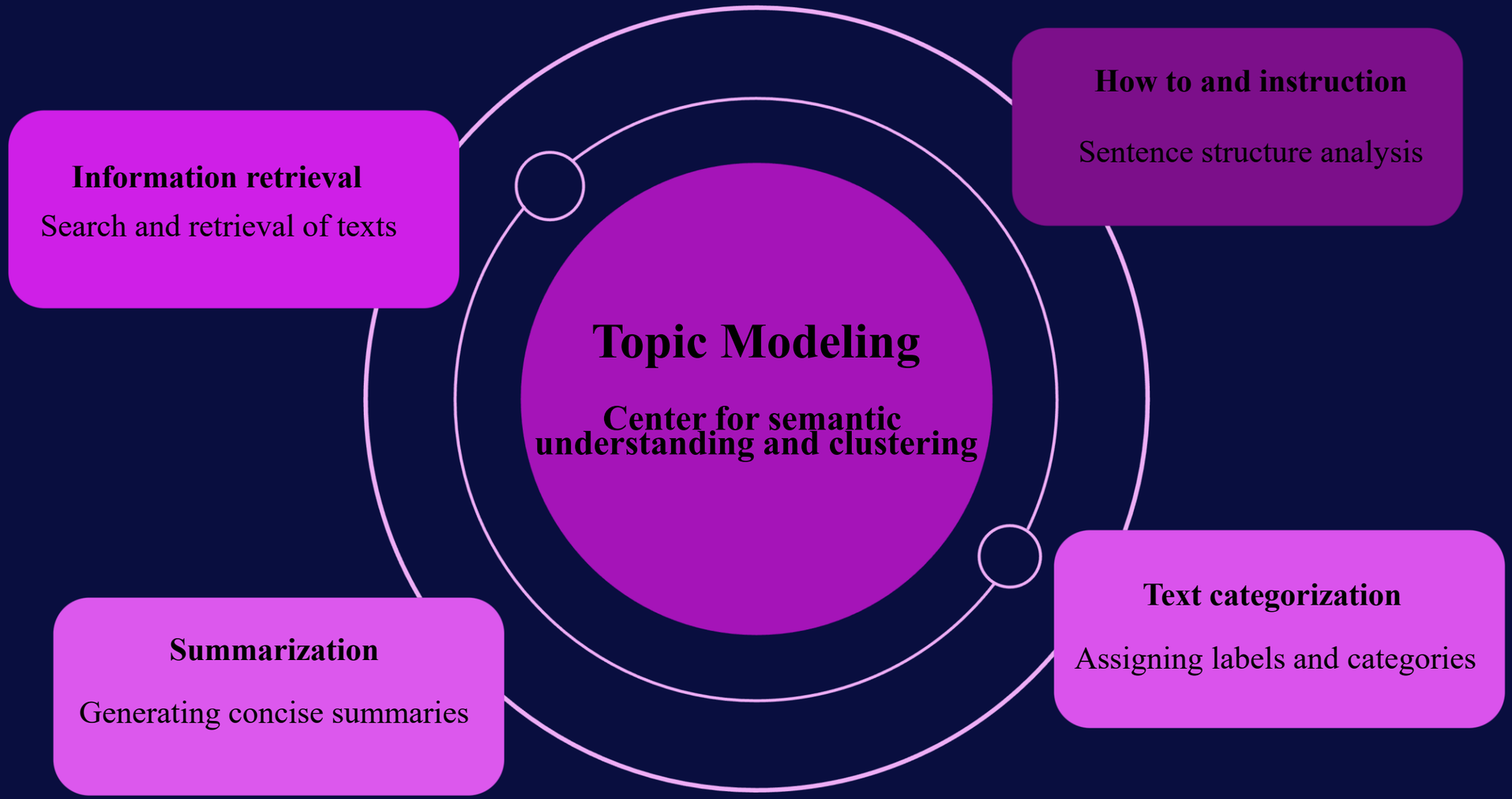
- **Classification and prioritization of security incidents**

  A lengthy security report can be quickly decomposed into its main topics and forwarded to the relevant unit.

Relationship with other NLP domains

- **Text Mining:** The core operation for discovering textual patterns.
- **Summarization:** Aids in identifying main topics for summarization.
- **Text Classification:** Can be used as a preprocessing stage.
- **Recommender Systems:** Creates user topic profiles.

The position of Topic Modeling in NLP



A look at leading models

This field has witnessed the development of advanced models, each offering distinct approaches to topic discovery.

BER Topic

A pipeline consisting of a Transformer Encoder (BERT), dimensionality reduction (UMAP), clustering (HDBSCAN), and topic representation (c-TF-IDF). Highly efficient and widely used.

Neural Topic Models + Transformer

Models such as CTM (Contextualized Topic Models) and Top2Vec, which combine the power of neural networks and transformers.

Conceptual Summary

Transformer-based Topic Modeling represents a new generation of topic discovery models. By leveraging the semantic understanding power of transformers, it enables **accurate, scalable, and interpretable extraction of topics** from complex textual data.

Literature review

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 244 (2024) 121–132

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on AI in Computational Linguistics

Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms

Melkamu Abay Mersha[a], Mesay Gemeda yigezu *[b], Jugal Kalita[a]

[a]College of Engineering and Applied Science, University of Colorado Colorado Springs (UCCS), Colorado Springs, USA
[b]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico city, Mexico

Abstract

Topic modeling is a powerful technique to discover hidden topics and patterns within a collection of documents without prior knowledge. Traditional topic modeling and clustering-based techniques encounter challenges in capturing contextual semantic information. This study introduces an innovative end-to-end semantic-driven topic modeling technique for the topic extraction process, utilizing advanced word and document embeddings combined with a powerful clustering algorithm. This semantic-driven approach represents a significant advancement in topic modeling methodologies. It leverages contextual semantic information to extract coherent and meaningful topics. Specifically, our model generates document embeddings using pre-trained transformer-based language models, reduces the dimensions of the embeddings, clusters the embeddings based on semantic similarity, and generates coherent topics for each cluster. Compared to ChatGPT and traditional topic modeling algorithms, our model provides more coherent and meaningful topics.

Peer-review under responsibility of the scientific committee of the 6th International Conference on AI in Computational Linguistics, ACLing 2024

Keywords: Topic Modeling; Semantic; Cluster; Transformer-Based Embeddings; Transformer; Topic Extraction; Semantic-Driven; Deep Learning.

CWTM: Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling

Zheng Fang[1], **Yulan He**[1,2,3] **and Rob Procter**[1,3]
[1]Department of Computer Science, University of Warwick, UK,
[2]Department of Informatics, King's College London, UK,
[3]The Alan Turing Institute, UK
{Z.Fang.4, Rob.Procter}@warwick.ac.uk
yulan.he@kcl.ac.uk

Abstract

Most existing topic models rely on bag-of-words (BOW) representation, which limits their ability to capture word order information and leads to challenges with out-of-vocabulary (OOV) words in new documents. Contextualized word embeddings, however, show superiority in word sense disambiguation and effectively address the OOV issue. In this work, we introduce a novel neural topic model called the Contextlized Word Topic Model (CWTM), which integrates contextualized word embeddings from BERT. The model is capable of learning the topic vector of a document without BOW information. In addition, it can also derive the topic vectors for individual words within a document based on their contextualized word embeddings. Experiments across various datasets show that CWTM generates more coherent and meaningful topics compared to existing topic models, while also accommodating unseen words in newly encountered documents.

Keywords: Topic Modeling, BERT, Contextualized Word Embeddings

1. Introduction

Topic modeling has been widely used to explore latent themes within vast document collections, where each document is modeled as a mixture of topics, and a topic is represented by a list of words sorted by their co-occurrence association strength with the topic. Most existing topic models rely on bag-of-words (BOW) representation, which limits their ability to capture word order information and leads to challenges with out-of-vocabulary (OOV) words in new documents. Contextualized word embeddings, however, show superiority in word sense disambiguation and effectively address the OOV issue. In this work, we introduce a novel neural topic model called the Contextlized Word Topic Model (CWTM), which integrates contextualized word embeddings from BERT. The model is capable of learning the topic vector of a document without BOW information. In addition, it can also derive the topic vectors for individual words within a document based on their contextualized word embeddings. Experiments across various datasets show that CWTM generates more coherent and meaningful topics compared to existing topic models, while also accommodating unseen words in newly encountered documents.

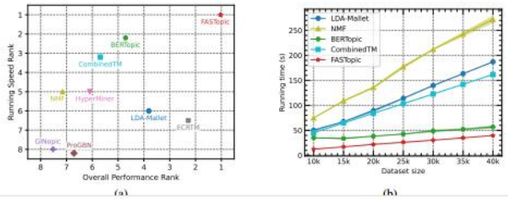
natural language processing (NLP) research has entered a new era. By capturing the surrounding context of each word, these models generate distinct contextualized word embeddings for each occurrence of a word in text. Contextualized word embeddings are superior to static word embeddings for word sense disambiguation, as they grasp the contextual nuances surrounding word

FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model

Xiaobao Wu[1]* **Thong Nguyen**[2] **Delvin Ce Zhang**[3] **William Yang Wang**[4] **Anh Tuan Luu**[1]*
[1]Nanyang Technological University [2]National University of Singapore
[3]The Pennsylvania State University [4]University of California, Santa Barbara
xiaobao002@e.ntu.edu.sg e0998147@u.nus.edu delvin.ce.zhang@gmail.com
william@cs.ucsb.edu anhtuan.luu@ntu.edu.sg

Abstract

Topic models have been evolving rapidly over the years, from conventional to recent neural models. However, existing topic models generally struggle with either effectiveness, efficiency, or stability, highly impeding their practical applications. In this paper, we propose FASTopic, a fast, adaptive, stable, and transferable topic model. FASTopic follows a new paradigm: Dual Semantic-relation Reconstruction (DSR). Instead of previous conventional, VAE-based, or clustering-based methods, DSR directly models the semantic relations among document embeddings from a pretrained Transformer and learnable topic and word embeddings. By reconstructing through these semantic relations, DSR discovers latent topics. This brings about a neat and efficient topic modeling framework. We further propose a novel Embedding Transport Plan (ETP) method. Rather than early straightforward approaches, ETP explicitly regularizes the semantic relations as optimal transport plans. This addresses the relation bias issue and thus leads to effective topic modeling. Extensive experiments on benchmark datasets demonstrate that our FASTopic shows superior effectiveness, efficiency, adaptivity, stability, and transferability, compared to state-of-the-art baselines across various scenarios.



Artificial Intelligence Review (2024) 57:18
https://doi.org/10.1007/s10462-023-10661-7

A survey on neural topic models: methods, applications, and challenges

Xiaobao Wu[1] · **Thong Nguyen**[2] · **Anh Tuan Luu**[1]

Accepted: 19 December 2023 / Published online: 25 January 2024

Abstract

Topic models have been prevalent for decades to discover latent topics and infer topic proportions of documents in an unsupervised fashion. They have been widely used in various applications like text analysis and context recommendation. Recently, the rise of neural networks has facilitated the emergence of a new research field—neural topic models (NTMs). Different from conventional topic models, NTMs directly optimize parameters without requiring model-specific derivations. This endows NTMs with better scalability and flexibility, resulting in significant research attention and plentiful new methods and applications. In this paper, we present a comprehensive survey on neural topic models concerning methods, applications, and challenges. Specifically, we systematically organize current NTM methods according to their network structures and introduce the NTMs for various scenarios like short texts and cross-lingual documents. We also discuss a wide range of popular applications built on NTMs. Finally, we highlight the challenges confronted by NTMs to inspire future research.

Keywords Neural topic model · Cross-lingual topic modeling · Dynamic topic modeling · Short text topic modeling · Variational AutoEncoder

BERTopic for Topic Modeling of Hindi Short Texts: A Comparative Study

Atharva Mutsaddi, Anvi Jamkhande, Aryan Thakre, Yashodhara Haribhakta
Department of Computer Science and Engineering, COEP Technological University
{atharvaam21, jamkhandeaa21, aryanst21, ybl}.comp@coeptech.ac.in

Abstract

As short text data in native languages like Hindi increasingly appear in modern media, robust methods for topic modeling on such data have gained importance. This study investigates the performance of BERTopic in modeling Hindi short texts, an area that has been under-explored in existing research. Using contextual embeddings, BERTopic can capture semantic relationships in data, making it potentially more effective than traditional models, especially for short and diverse texts. We evaluate BERTopic using 6 different document embedding models and compare its performance against 8 established topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Indexing (LSI), Additive Regularization of Topic Models (ARTM), Probabilistic Latent Semantic Analysis (PLSA), Embedded Topic Model (ETM), Combined Topic Model (CTM), and Top2Vec. The models are assessed using coherence scores across a range of topic counts. Our results reveal that BERTopic consistently outperforms other models in capturing coherent topics from short Hindi texts.

1 Introduction

Topic modeling is a widely-used technique in text mining that identifies underlying themes within textual data. BERTopic, a newer model in this field, has demonstrated its effectiveness by using pre-trained document embedding models and unsupervised clustering algorithms to form topic groups with high semantic coherence (Grootendorst, 2022). Unlike traditional models, BERTopic's use of embeddings allows it to capture semantic relationships in data, making it potentially more effective than traditional models, especially for short and diverse texts. We evaluate BERTopic using 6 different document embedding models and compare its performance against 8 established topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Indexing (LSI), Additive Regularization of Topic Models (ARTM), Probabilistic Latent Semantic Analysis (PLSA), Embedded Topic Model (ETM), Combined Topic Model (CTM), and Top2Vec. The models are assessed using coherence scores across a range of topic counts. Our results reveal that BERTopic consistently outperforms other models in capturing coherent topics from short Hindi texts.

largely focused on traditional methods that rely on probabilistic frameworks and matrix factorisation, which often overlook natural language semantics (Ray et al., 2019). Also, these studies primarily focus on long text documents, leaving a gap in the exploration of short Hindi texts, which are increasingly common in todays digital landscape.

Topic modeling in Hindi faces several unique challenges. Hindi does not use capitalisation to differentiate proper nouns from other word forms, complicating named entity recognition. Additionally, the lack of standardised spelling leads to multiple variations of the same word (Figure 1), creating ambiguity. Hindi also often employs repetitive expressions for emphasis, which can affect tokenization and cross-language natural language processing tasks (Ray et al., 2019).

This study aims to demonstrate that traditional topic models often fall short in capturing the semantic meaning of Hindi text due to these inherent challenges and struggle with the nuances of short texts where semantic meaning is more compressed and context-dependent. The contributions of this paper are as follows:

- Conducting a comprehensive comparison of BERTopic's performance across several aspects:
  - Evaluating BERTopic using different sentence transformer models such as `HindSBERT-STS` (Joshi et al., 2022), `XLM-R` (XLM-RoBERTa) (Conneau et al., 2020), `IndicBERT` (Kakwani et al., 2020), and `mBERT` (Multilingual BERT) (Devlin et al., 2018), and analysing results using coherence met-

Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms

Mersha, M. A., Yigezu, M. G., & Kalita, J. (2024). Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms. *Procedia Computer Science, 244*, 121–132.

Main objective of the article:

This article addresses the key limitation of traditional Topic Modeling methods: the inability to comprehend context-dependent meaning. Methods such as LDA operate solely based on word co-occurrence and fail to capture conceptual semantics. Even certain Transformer-based models do not fully utilize the linguistic understanding power of these models in the final stage of topic extraction.

Proposed Method Architecture:

A. Unsupervised Pipeline:
B. Document Embedding
C. Dimension Reduction
D. Document Clustering
E. Topic Extraction

Key Results:

The proposed model demonstrated significantly better performance across all metrics for topic coherence compared to BERTopic and traditional models.

The model's output generated interpretable and meaningful topics.

Weaknesses or Limitations:

- High computational cost
- Sensitivity to parameters
- Failure to identify hidden subtopics: Weakness in discovering topics embedded within the text.
- Limited scope of evaluation: Other metrics, such as topic diversity, were less explored, with the main focus being on coherence.

Strengths of the Method:

- **Semantic-centric:** Focuses on the **concept** of words rather than their **repetition**.
- **Modular design**
- **Unsupervised**
- **Automatic noise removal and outlier data management**
- **Controllability:** Users can adjust the **similarity threshold** to merge or separate topics.

CWTM: Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling

Zheng Fang[1], **Yulan He**[1,2,3] **and Rob Procter**[1,3] [1]Department of Computer Science, University of Warwick, UK, [2]Department of Informatics, King's College London, UK, [3]The Alan Turing Institute, UK

Main objective of the article:

This article introduces a new Topic Modeling model called CWTM, which leverages Contextualized Word Embeddings from the BERT language model to overcome the limitations of classic Bag-of-Words (BOW)-based Topic Modeling. The primary goal is to improve topic coherence, manage the vocabulary, and extract more meaningful topics without relying on BOW representation.

Strengths of the Method:

- Independence from BOW
- Utilization of Contextualized Embeddings
- Robustness against novel vocabulary
- Higher topic coherence and diversity

Weaknesses or Limitations:

Dependence on a pre-trained model

Complexity of parameter tuning

Limited interpretability

Dependence on the quality of BERT embeddings

Key Results:

By eliminating reliance on BOW and directly utilizing **Contextualized Word Embeddings**, the model not only generates **more coherent and diverse topics**, but also demonstrates **greater robustness** when faced with novel data and unfamiliar vocabulary. This approach can also lead to **improved performance in practical applications**.

Proposed Method Architecture:

The CWTM model consists of 4 stages:

- Data Preparation
- Extracting Contextualized Word Embeddings
- Learning Topic Vectors for Each Word
- Constructing Document Topic Vector
- Learning Document Representation Vector
- Reconstruction and Distribution Matching
- Multi-Objective Training
- Topic Extraction

FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model

Xiaobao Wu[1], Thong Nguyen[2], Delvin Ce Zhang[3], William Yang Wang[4], Anh Tuan Luu 2024

Main Objective of the Article:

To propose a novel topic model called **FASTopic**, which consists of two core components: **Dual Semantic-relation Reconstruction** and **Embedding Transport Plan**. Instead of relying on classical methods, it directly models the semantic relationships between documents, topics, and words—aiming to uncover hidden topics in texts with enhanced **speed, adaptability, stability, transferability, and efficiency**.

Strengths of the method:

- Very high execution speed
- Better performance in Topic Coherence and Topic Diversity metrics
- High stability
- Simple architecture

Weaknesses or limitations:

- Dependence on a pre-trained model
- Limitation in processing long documents
- High memory consumption
- Lack of support for multilingual documents
- Initial parameter tuning

Proposed method architecture:

A. Document Embedding Extraction
B. Random Initialization of Topic and Word Embeddings
C. Dual Semantic-Relation Modeling via Optimal Transport
D. Document Reconstruction via Semantic Relations
E. Joint Optimization
F. Inference for New Documents

Key results:

The article demonstrates that FASTopic significantly outperforms leading models across all evaluation metrics—including topic quality, document-topic distribution quality, execution speed, stability, transferability, and overall performance—while delivering an optimal balance between efficiency, speed, and stability.

BERTopic for Topic Modeling of Hindi Short Texts: A Comparative Study

Atharva Mutsaddi, Anvi Jamkhande, Aryan Thakre, Yashodhara Haribhakta
Department of Computer Science and Engineering, COEP Technological University {atharvaam21, jamkhandeaa21, aryanst21, ybl}.comp@coeptech.ac.in

Main objective of the article:

This article evaluates the effectiveness of the BERTopic model for topic modeling in Hindi short texts and compares it with other classical and modern methods to demonstrate its superiority in addressing specific linguistic challenges of Hindi, such as the absence of capital letters, non-standard spelling, and repetitive emphatic phrases.

Weaknesses or limitations:

- Data limitation and the use of a single dataset
- Topic bias in the dataset
- Exclusive focus on short texts and lack of testing the model on longer documents
- Lack of multilingual support

Strengths of the method:

- Focus on low-resource languages
- Comprehensive comparison
- Qualitative analysis
- Multiple embedding evaluation
- Qualitative cluster analysis

Proposed method architecture:

A. Document embedding extraction
B. Dimensionality reduction
C. Clustering
D. Topic extraction
E. Evaluation

Key results:

BERTopic with **mBERT-Uncased embeddings** consistently outperformed all compared models in topic modeling for Hindi short texts and was able to extract more coherent and meaningful topics even in the presence of specific linguistic challenges of Hindi.

A survey on neural topic models: methods, applications, and challenges

Xiaobao Wu[1] · Thong Nguyen[2] · Anh Tuan Luu[1]

Accepted: 19 December 2023 / Published online: 25 January 2024

Main objective of the article:

The article provides a comprehensive and up-to-date survey of Neural Topic Models (NTMs). It begins by comparing NTMs with classical topic models such as LDA, emphasizing the advantages of NTMs in terms of flexibility, scalability, and the absence of model-specific inference requirements.

Strengths of NTMs:

- High flexibility
- Scalability
- Integration with advanced neural architectures
- Effectiveness in complex scenarios
- Gradient-based learning

Weaknesses or limitations of NTMs:

- Low quality of generated topics
- Sensitivity to hyperparameters
- Lack of standard and reliable evaluation metrics
- Data volume dependency – better performance with large datasets

Categories reviewed in the article:

Based on application scenarios:

- ✓ Hierarchical NTMs
- ✓ Short Text NTMs
- ✓ Cross-lingual NTMs
- ✓ Dynamic NTMs
- ✓ Correlated NTMs
- ✓ Lifelong NTMs

Categories reviewed in the article:

Based on different neural architectures:

- ✓ NTMs with Various Priors
- ✓ NTMs with Embeddings
- ✓ NTMs with Metadata
- ✓ NTMs with Graph Neural Networks
- ✓ NTMs with Generative Adversarial Networks
- ✓ NTMs with Pre-trained Language Models
- ✓ NTMs with Contrastive Learning
- ✓ NTMs with Reinforcement Learning
- ✓ Topic Discovery by Clustering

Key results:

By presenting a structured survey, the article demonstrates that NTMs have made significant advancements in flexibility, scalability, and integration with neural architectures. The article suggests that future efforts should focus on improving topic quality, standardizing evaluation, and reducing sensitivity to parameters.

Research Directions: Evolutions and the Future

Research in Transformer-based topic modeling is pursued along five main streams, each focusing on a specific aspect of this technology.

Stream 5

Human-in-the-Loop and Interactive Topic Modeling (Integration with human feedback)

Stream 4

Dynamic and Temporal Topic Modeling (Modeling topic evolution over time)

Stream 3

End-to-End Transformer Topic Models – Unified topic modeling

Stream ۲

Neural Topic Models with Contextual Embeddings (Combining NTMs with BERT)

Stream 1

Replacing Bag-of-Words with semantic embeddings (BERTopic, Top2Vec)

Challenges and Research Gaps

Despite significant progress, this field still faces challenges that shape the trajectory of future research.

Research Gaps

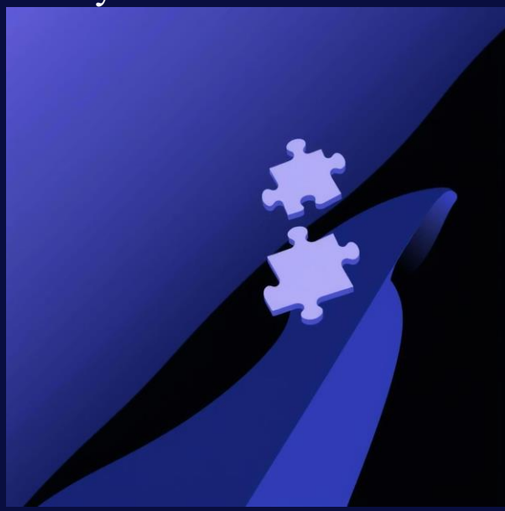
- Lack of a Standard Benchmark: Shortage of shared datasets, evaluation metrics, and application scenarios.
- Poor Evaluation in Real-World Applications: Excessive focus on Topic Coherence and insufficient attention to business value and decision impact.

Fundamental Challenges

- Absence of a formal definition of a "topic:" Makes fair comparison of models difficult.
- Conflict between semantic quality and interpretability: Models with high semantic quality sometimes have lower interpretability.

Technical Challenges

- Scalability for organizational data: Processing millions of documents under GPU constraints.
- Topic instability: Re-running the model may lead to different topics.
- Dependency on Clustering: The choice of clustering algorithm significantly impacts the output.



The Future of Topic Modeling: Integration with LLMs and Intelligent Systems

Potential pathways for future research in this field are moving toward more advanced applications and deeper integration with modern technologies.

Topic Modeling as Semantic Infrastructure

Usage as an intermediate layer in intelligent systems (Search, Recommendation, Knowledge Graph).

Integration with LLMs

Using LLMs for automatic labeling, merging/separating topics, and qualitative evaluation. Example: BERTopic + GPT.

Dynamic and Real-time Topic Modeling

Streaming Topic Modeling, Online Learning, and Drift-Aware Models for social media monitoring and crisis management.

Explainable & Policy-Aware Topic Models

Generating reasons for topic formation and evolution paths, suitable for policy-making and macro-level decision-making.

Domain-Specific Topic Modeling

Developing legal, financial, healthcare, and telecom models using Fine-Tuned Transformers.

> Transformer-based Topic Modeling, as the new generation of topic discovery models, has successfully overcome the semantic understanding limitations of classical models. The future trajectory of this field is moving toward integration with LLMs, dynamic models, decision support systems, and domain-specific applications.