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

Student: Hengameh Shah Ali
40307604

Problem Definition

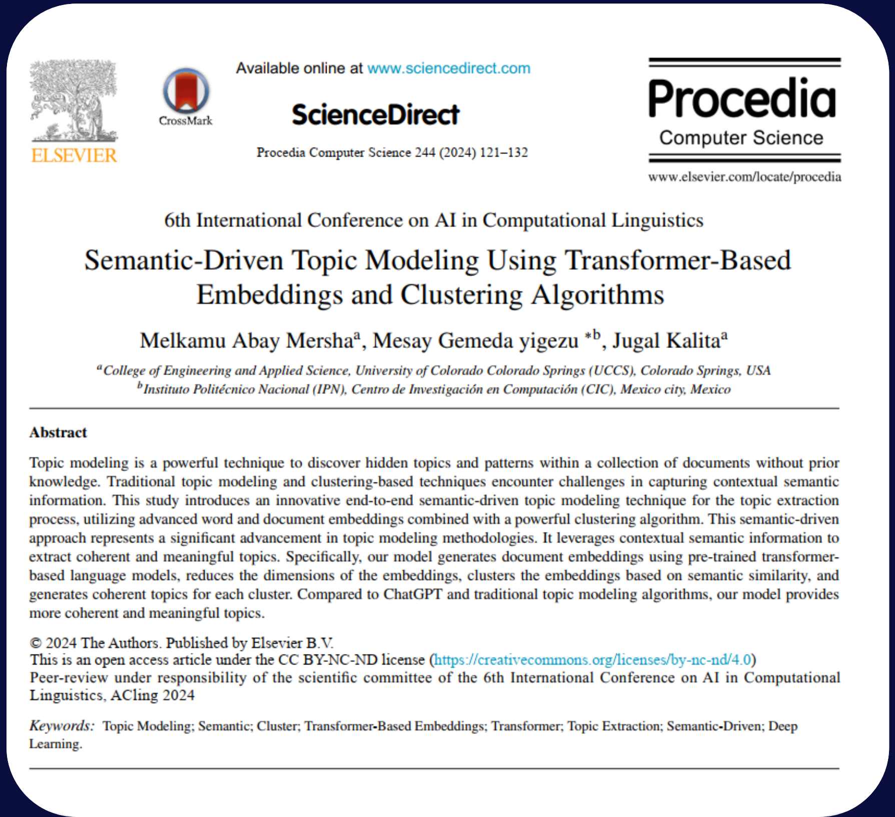
The subject of the paper is a powerful technique for discovering hidden topics and semantic patterns in a collection of text documents without requiring prior knowledge. Classical methods such as LDA, NMF, and LSA are based on the Bag-of-Words model, which ignores contextual meanings and relationships between words and merely counts word occurrences in the text without any understanding of the actual meaning of words within sentence context. This leads to two main issues:

- Lack of understanding of word meanings in sentences
- Generation of low-quality and incoherent topics

New language models such as Transformers understand complex semantic relationships, but many topic modeling techniques do not utilize this capability, resulting in the extraction of inconsistent, inaccurate, and meaningless topics.

Thus, the main problem addressed in the paper is:

How can topics be extracted based on meaning rather than just word frequency?



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.

- **Models Used:** all-MiniLM-L6-v2 and all-mpnet-base-v2

Implementation Method

❑ Document Embedding

Method Used: Sentence-BERT (SBERT)

Process: Text documents are converted into meaningful numerical vectors (embeddings).

Details: Each sentence receives a fixed-length vector representation using pre-trained transformer models.

❑ Dimensionality Reduction

Why Needed? High-dimensional vectors make clustering difficult (curse of dimensionality).

Method Used: UMAP (Uniform Manifold Approximation and Projection)

Details: UMAP parameters (number of neighbors, minimum distance) are tuned to preserve both global and local data structures.

❑ Document Clustering

Algorithm Used: HDBSCAN (Hierarchical Density-Based Spatial Clustering)

Why HDBSCAN?

- Finds clusters of varying densities
- Robust to noise
- Identifies outliers automatically
- No need to predefine the number of clusters

Clustering Basis: Semantic similarity of reduced document embeddings.

Implementation Method

❑ Topic Extraction

Process for Each Cluster:

- Vocabulary Construction: Words from each sentence are extracted and mapped to their embedding vectors.
- Irrelevant Word Removal: Words without semantic contribution to sentences are eliminated.
- Average Similarity Calculation: For each word, the average cosine similarity with all sentences in the cluster is computed .

$$\text{avg\_sim}(w_{i)=}\frac{1}{N}\sum_{j=1}^{N}\text{cosine}(\,w_i, s_j)$$

- Top Word Selection: The *k* words with the highest similarity scores are selected as the cluster's topic.
- Topic Merging: Topics with high similarity are merged to control the final number of topics.

❑ Topic Refinement and Evaluation

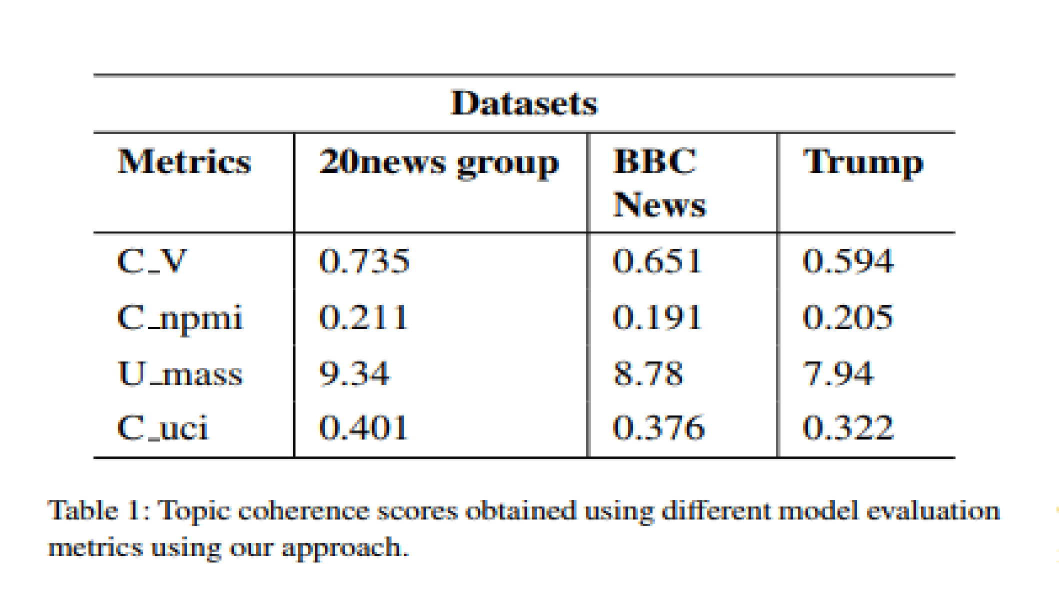
- Use metrics such as CV, Cnpmi, UMass, and Cuci to evaluate topic coherence.
- Adjust hyperparameters (e.g., minimum cluster size in HDBSCAN, k in topic words) based on validation results.
- Employ OCTIS toolkit for standardized comparison with baseline models (LDA, BERTopic, CTM, etc.).

Main Results of the Paper

Topic Coherence Evaluation

The proposed model achieved the highest topic coherence scores across all three datasets (20NewsGroups, BBC News, Trump's Tweets) compared to all baseline models. It attained notable and superior scores across all four evaluation metrics (C_V, C_npmi, U_Mass, and C_uci), demonstrating its consistent performance and effectiveness in extracting meaningful topics.(Table۱)

The best performance was observed on the **20NewsGroups dataset**, indicating that the model performs optimally on structured and long-form data (such as news articles). At the same time, the results demonstrate that the model also performs adequately on short texts like tweets, achieving a reasonable and competitive score.

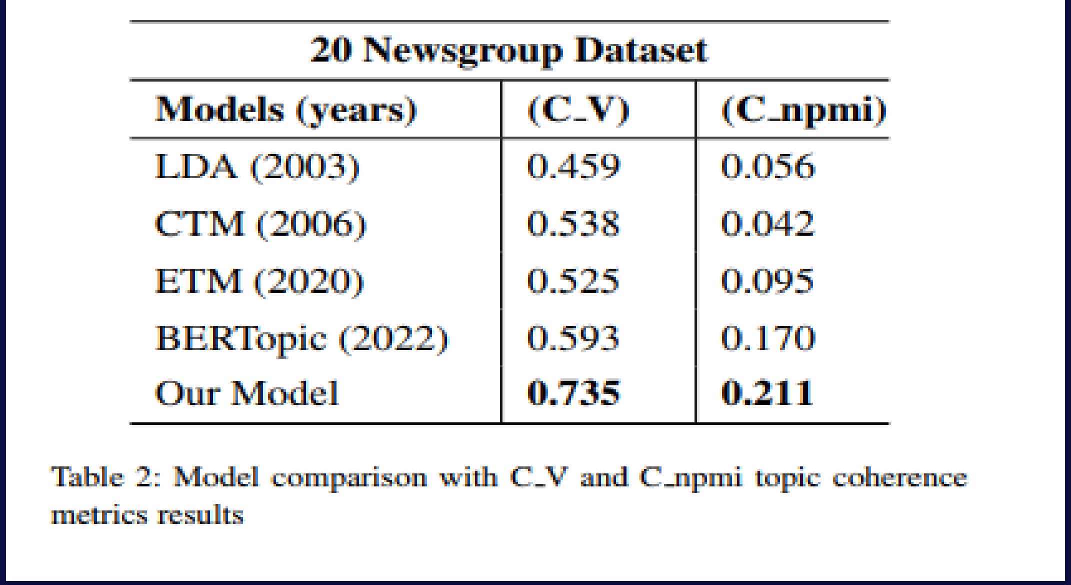


	Datasets		
Metrics	20news group	BBC News	Trump
C_V	0.735	0.651	0.594
C_npmi	0.211	0.191	0.205
U_mass	9.34	8.78	7.94
C_uci	0.401	0.376	0.322

Table 1: Topic coherence scores obtained using different model evaluation metrics using our approach.

Comparison with Traditional and Embedding-Based Models

The proposed model significantly outperformed both classical and modern models.(Table2)



20 Newsgroup Dataset		
Models (years)	(C_V)	(C_npmi)
LDA (2003)	0.459	0.056
CTM (2006)	0.538	0.042
ETM (2020)	0.525	0.095
BERTopic (2022)	0.593	0.170
Our Model	**0.735**	**0.211**

Table 2: Model comparison with C_V and C_npmi topic coherence metrics results

Comparison with ChatGPT:

ChatGPT was not suitable for large datasets and required chunking, which led to:

- Loss of latent themes
- Inability to process sequential data effectively
- Lack of built-in evaluation metrics

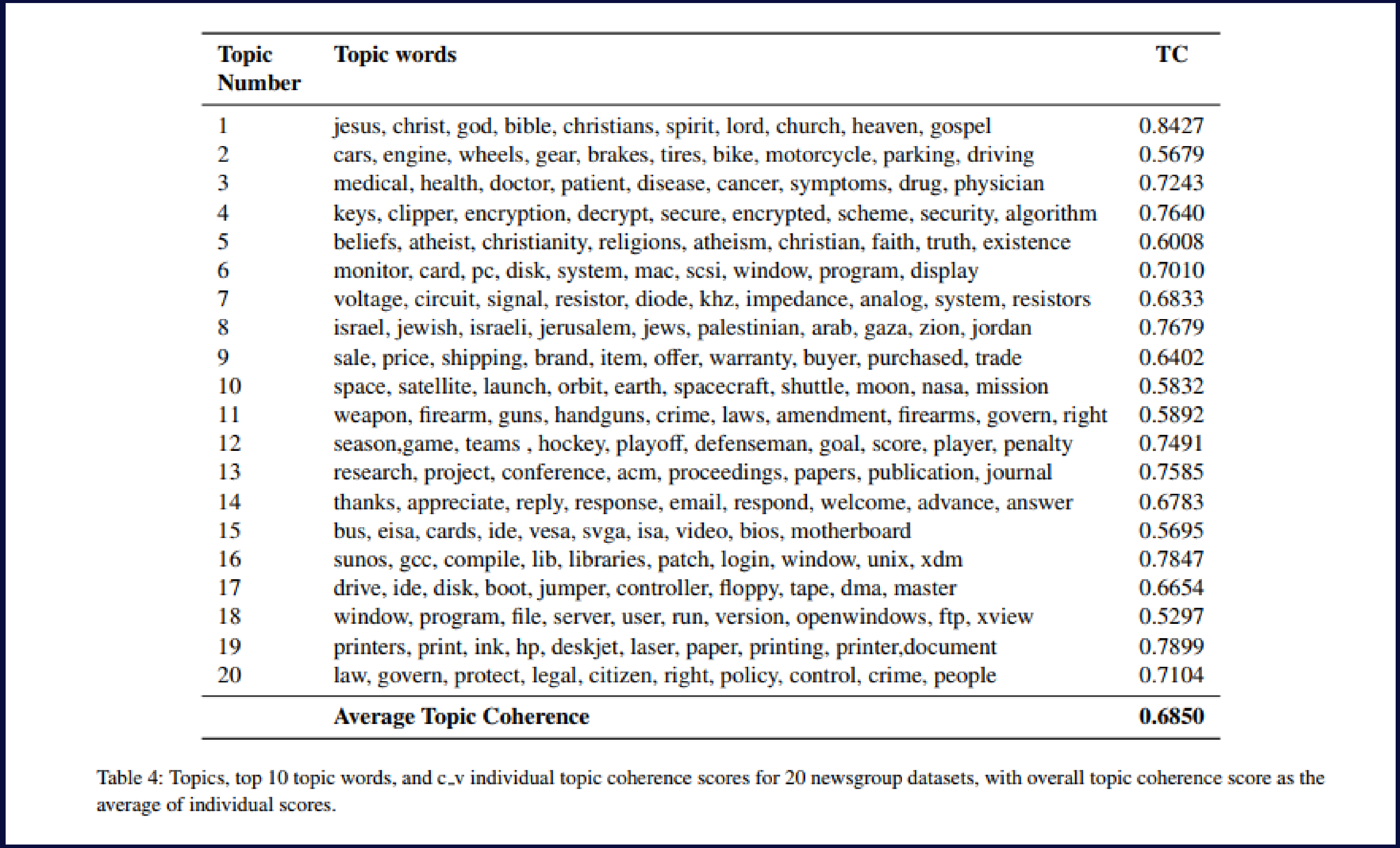
The proposed model performed better than ChatGPT in terms of **scalability, reliability, and topic quality**.

Sample Extracted Topics

For the **20NewsGroups** dataset, 20 meaningful topics were extracted, including:

- Religion: jesus, christ, god, bible, christians, ...
- Technology: monitor, card, pc, disk, system, ...
- Medical: medical, health, doctor, patient, disease, ...
- Sports: season, game, teams, hockey, playoff, ...
- Politics: law, govern, protect, legal, citizen, ...

Average topic coherence for these topics: **0.685**



Topic Number	Topic words	TC
1	jesus, christ, god, bible, christians, spirit, lord, church, heaven, gospel	0.8427
2	cars, engine, wheels, gear, brakes, tires, bike, motorcycle, parking, driving	0.5679
3	medical, health, doctor, patient, disease, cancer, symptoms, drug, physician	0.7243
4	keys, clipper, encryption, decrypt, secure, encrypted, scheme, security, algorithm	0.7640
5	beliefs, atheist, christianity, religions, atheism, christian, faith, truth, existence	0.6008
6	monitor, card, pc, disk, system, mac, scsi, window, program, display	0.7010
7	voltage, circuit, signal, resistor, diode, khz, impedance, analog, system, resistors	0.6833
8	israel, jewish, israeli, jerusalem, jews, palestinian, arab, gaza, zion, jordan	0.7679
9	sale, price, shipping, brand, item, offer, warranty, buyer, purchased, trade	0.6402
10	space, satellite, launch, orbit, earth, spacecraft, shuttle, moon, nasa, mission	0.5832
11	weapon, firearm, guns, handguns, crime, laws, amendment, firearms, govern, right	0.5892
12	season,game, teams , hockey, playoff, defenseman, goal, score, player, penalty	0.7491
13	research, project, conference, acm, proceedings, papers, publication, journal	0.7585
14	thanks, appreciate, reply, response, email, respond, welcome, advance, answer	0.6783
15	bus, eisa, cards, ide, vesa, svga, isa, video, bios, motherboard	0.5695
16	sunos, gcc, compile, lib, libraries, patch, login, window, unix, xdm	0.7847
17	drive, ide, disk, boot, jumper, controller, floppy, tape, dma, master	0.6654
18	window, program, file, server, user, run, version, openwindows, ftp, xview	0.5297
19	printers, print, ink, hp, deskjet, laser, paper, printing, printer,document	0.7899
20	law, govern, protect, legal, citizen, right, policy, control, crime, people	0.7104
	Average Topic Coherence	**0.6850**

Table 4: Topics, top 10 topic words, and c_v individual topic coherence scores for 20 newsgroup datasets, with overall topic coherence score as the average of individual scores.

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.

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).

❑ **Content Recommender System:** Recommending articles, news, and content based on user interests

Extract topics from the content consumed by the user

Match with topics of new content

Suggest content with similar topics

❑ **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.