Mining the Past: A Comparative Study of Classical and Neural Topic Models on Historical Newspaper Archives

Keerthana Murugaraj[1], Salima Lamsiyah[1], Marten During[2], Martin Theobald[1]
[1]Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg
[2] Centre for Contemporary & Digitial History (C[2]DH), University of Luxembourg
Correspondence: keerthana.murugaraj@uni.lu

Abstract

Analyzing historical discourse in large-scale newspaper archives requires scalable and interpretable methods to uncover hidden themes. This study systematically evaluates topic modeling approaches for newspaper articles from 1955 to 2018, comparing probabilistic *LDA*, matrix factorization *NMF*, and neural-based models such as *Top2Vec* and *BERTopic* across various preprocessing strategies. We benchmark these methods on topic coherence, diversity, scalability, and interpretability. While LDA is commonly used in historical text analysis, our findings demonstrate that BERTopic, leveraging contextual embeddings, consistently outperforms classical models in all tested aspects, making it a more robust choice for large-scale textual corpora. Additionally, we highlight the trade-offs between preprocessing strategies and model performance, emphasizing the importance of tailored pipeline design. These insights advance the field of historical NLP, offering concrete guidance for historians and computational social scientists in selecting the most effective topic-modeling approach for analyzing digitized archives. Our code will be publicly available on GitHub.

1 Introduction

Digitized newspapers have become widely used in recent years, providing convenient access to extensive historical records. Online platforms further support historians in efficiently identifying and analyzing primary and secondary sources (Allen and Sieczkiewicz, 2010). However, the vast amount of documents and information available presents a challenge for historians in terms of study, analysis, and interpretation. To address these challenges, Natural Language Processing (NLP) methods are frequently employed to streamline the process. In our recent work, we present a novel approach for both extractive and abstractive summarization of historical texts (Lamsiyah et al., 2023; Murugaraj et al., 2025). In this paper, we focus on Topic Modeling (TM) methods to automatically extract themes from historical newspaper archives, reducing the time historians would otherwise spend on manually categorizing and analyzing these contents.

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) laid the foundation for TM. Building on this, the probabilistic framework known as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) was introduced. However, the development of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) represents a significant turning point in the field, providing a more sophisticated and effective probabilistic approach for uncovering latent topics within large-scale text corpora. Another widely used technique is Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999), which employs matrix factorization technique by decomposing a term-document matrix into two low-dimensional, non-negative matrices representing words and documents.

Building on these foundational methods, many new approaches have emerged in recent years. The introduction of the Transformer architecture (Vaswani et al., 2017) revolutionized many NLP aspects and paved the way for the development of advanced neural-network models. Since then, traditional TM techniques have been enhanced by neural-based methods that leverage contextual embeddings. Among these, two widely used approaches are Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022). These models demonstrate promising performance in capturing contextual meaning and intricate patterns within textual data, significantly outperforming conventional methods. While LDA and NMF have been widely applied across various fields, including historical research, neural topic models still remain underutilized in this domain.

Egger and Yu (2022) compared four topic mod-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 452–463

els on Twitter posts, which are short texts, using qualitative evaluation. However, these findings cannot be directly applied to newspaper articles, as they often cover multiple topics within the same document. Given the structured and in-depth nature of news articles, it is crucial to evaluate topic models, specifically in this context. To address the challenge of selecting the best topic-modeling approach for historical newspaper articles, we conduct a comprehensive empirical evaluation of classical and neural topic models on a large historical newspaper dataset. The main contributions of our work are as follows:

- We highlight the crucial role of preprocessing, showing that extensive preprocessing improves topic coherence and diversity.
- We show that embedding models with extended input lengths improve topic quality, while smaller models require careful chunking and aggregation strategies for comparable performance.
- We show that BERTopic outperforms traditional (LDA, NMF) and neural (Top2Vec) models in extracting key topics from historical news archives, with stable performance across all data subsets, highlighting its reliability and adaptability for historical topic modeling.

By systematically analyzing various preprocessing methods, different embedding models, and model performance, we offer tailored recommendations for analyzing historical archives. To the best of our knowledge, this study is the first comprehensive comparison of four topic-modeling methods specifically applied to a large-scale historical news archive.

2 Related Works

This section reviews historical topic modeling, existing approaches, and future directions.

Classical Topic Modeling Methods, such as LDA and NMF, have been widely used in the historical domain for topic detection. LDA (Blei et al., 2003) is a probabilistic model that represents documents as topic mixtures and topics as word distributions, using inference algorithms to estimate these topic distributions. NMF (Lee and Seung, 1999) is based on matrix decomposition, where the document-term matrix is factorized into two non-negative matrices representing the topics and their corresponding word distributions.

Several works have employed these classical models in historical research. Hall et al. (2008) conducted a study to explore the development of ideas in the field of Computational Linguistics over time by applying LDA to the ACL Anthology, covering the years 1978 to 2006. Yang et al. (2011) leveraged the LDA topic model on the collection of digitized historical newspapers published in Texas from 1829 to 2008. A very interesting study by Fridlund and Brauer (2013) provides a comprehensive overview of the history and application of TM within digital humanities, particularly in digital history from 2006 until 2012. Only 23 historical TM studies were found during 2006–2012, and the majority were conducted to explore the topic methods users and its usage rather than using it for solving independent historical questions. Another study by Gavin and Gidal (2016) conducted LDA-based TM to study the industrial and environmental history in Scotland. Ambrosino et al. (2018) also experimented with LDA to the large archives of economic articles. Zamiraylova and Mitrofanova (2020) study leverages the NMF algorithm to automatically identify and analyze dynamic topics within a corpus of Russian short stories from the first third of the 20th century, providing a deeper understanding of the thematic evolution in the Russian literature. The recent studies in the historical domain continue to strongly rely on LDA and NMF (Oiva, 2020; Marjanen et al., 2020; Maltseva et al., 2021; Bodrunova, 2021; Uban et al., 2021; Grant et al., 2021; Gryaznova and Kirina, 2021; Lin and Peng, 2022; Baklāne and Saulespurēns, 2022; Bourgeois et al., 2022; Grassia et al., 2022; Karamouzi et al., 2024; Chappelle et al., 2024).

Neural Topic Modeling Methods have gained popularity for capturing complex text relationships using deep learning. Recent TM methods, such as Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022), leverage neural embeddings and clustering techniques to improve topic discovery, offering greater flexibility and coherence compared to classical methods. Only very few studies have applied neural models in historical TM. Arseniev-Koehler et al. (2020) proposed Discourse Atom Topic Modeling (DATM), a novel method, that integrates probabilistic topic modeling with word embeddings applied to violent death narratives in the U.S. National Violent Death Reporting System, revealing nuanced themes and gender biases. Cvejoski et al. (2023) introduced the Neural Dynamic Focused Topic Model (NDF-TM), which

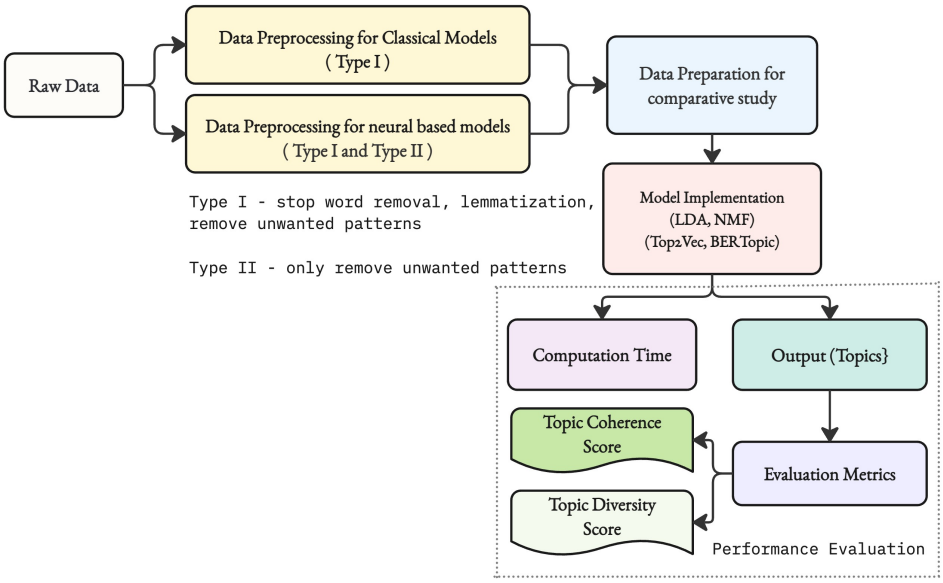

Figure 1: Methodology Workflow.

uses Bernoulli random variables and outperforms classical dynamic topic models in tracking topic evolution across the UN General Debates, NeurIPS papers, and ACL Anthology datasets. Martinelli et al. (2024) compared two neural topic models, Product-of-Experts LDA, and Embedded Topic Model, against LDA on a classical Latin corpus. Their evaluation found that neural models outperformed LDA in both quantitative metrics and expert qualitative assessments. Ginn and Hulden (2024) compared dynamic topic models on 1,350 Roman literature texts, finding that neural models aligned better with historical intuitions than classical models.

Shortcomings. Classical topic models fall short in capturing historical text semantics, while neural models provide richer, context-aware representations. Other domains have advanced by adopting neural topic models, which use deep learning techniques for more sophisticated and accurate topic representations. (Orr et al., 2024; Rajwal et al., 2024). The adoption of neural-based topic models in historical research remains limited. Only a handful of studies have ventured into using neural approaches so far, thus leaving a significant gap in the methodological toolkit available for historians. This lag indicates a pressing need for the historical domain to embrace and experiment with neural-based TM techniques. Motivated by this prevalent gap, we empirically analyzed classical and neural-based topic methods. Specifically, we picked two classical models, LDA and NMF, which are popularly used in the historical domain as baselines, and we compared them with the more recent Top2Vec and BERTopic neural-based models. We tested all methods on a large collection of more than 148,000 historical newspaper articles centered around the themes of "*nuclear power*" and "*nuclear safety*" to evaluate their performance.

3 Methodology

In this section, we outline the workflow used to conduct our study as presented in Figure 1.

Dataset Collection. The dataset was collected from historical archives[1], it spans nearly six decades of public and media narratives, segmented into four subsets: 1955–1970, 1971–1986, 1987–2002, 2003–2018. This segmentation provides a rich foundation for applying topic modeling to extract meaningful insights on societal, political, and other themes. Each document is assigned a unique identifier, ensuring precise referencing and tracking throughout our entire analysis.

Data Preprocessing. We created two distinct datasets through different preprocessing procedures, each specifically designed to support different topic-modeling approaches for analyzing historical newspaper archives. The *Type 1* dataset was prepared for classical models including lowercasing, stopword removal, filtering unwanted patterns (e.g., random IDs, alphanumeric sequences, special symbols), punctuation removal, and lemmatization for improved topic coherence. The *Type 2*

[1]Reference omitted due to double-blind reviewing.

Year	Type	#Docs	#Words	#Vocabulary	Min	Max	Avg. Length
1955–1970	1	49,217	14,842,929	292,188	7	1,408	302
	2	49,217	29,784,865	306,793	8	2,544	605
1971–1986	1	53,308	11,967,980	274,470	6	2,441	225
	2	53,308	23,434,484	281,786	8	2,758	440
1987–2002	1	32,459	7,699,135	201,723	5	2,183	237
	2	32,459	14,548,662	209,105	12	2,593	448
2003–2018	1	13,252	3,001,507	118,190	6	2,990	227
	2	13,252	5,334,551	126,019	10	3,968	403

Table 1: Key statistics for year-based document segments: preprocessing type, total documents, word count, vocabulary size, min/max tokens, and average document length.

dataset, designed for neural models, retained sentence boundaries and full stops to preserve the textual structure, while also removing unwanted patterns, symbols, and excessive whitespace. All text was converted to lowercase for uniformity. Both datasets were preprocessed using a combination of regular expressions and the spaCy[2] NLP libraries (the latter for stopword removal and lemmatization).

Data Preparation. Our initial Exploratory Data Analysis (EDA) aimed to understand the structure of the datasets and extract key statistical insights essential for topic modeling. This step was critical in assessing the distribution of documents across different time periods and evaluating the suitability of the dataset. We examined key characteristics of the two preprocessed dataset types, including the total number of documents, word count, vocabulary size, minimum and maximum token count per document, and average document length for each yearly segment. These insights validated the effectiveness of our preprocessing steps and highlighted potential challenges, such as variations in document length and vocabulary shifts over time.

The EDA results, summarized in Table 1, played a crucial role in guiding our selection and empirical comparison of topic modeling methods. Given the dataset characteristics, we selected four topic models—LDA, NMF, Top2Vec, and BERTopic—each suited to different structural properties. LDA and NMF, which rely on word co-occurrence patterns, are effective for structured corpora with stable vocabulary distributions but may struggle with short documents or datasets with significant topic overlap. In contrast, Top2Vec and BERTopic, which leverage embeddings, are better suited for handling multi-topic documents and capturing vocabulary shifts over time. Additionally, EDA ensured a fair comparison by identifying potential biases, such as imbalanced document lengths or topic sparsity, that could affect model evaluation. By aligning topic model selection with empirical dataset properties, EDA strengthens the interpretability and robustness of our comparative analysis.

Platform. We leveraged the recent OCTIS (Terragni et al., 2021a) toolkit for running models within its unified framework, which offers standardized procedures for evaluating topic-modeling algorithms. We prepared the data for all the methods according to its supported format.

Model	Model Params	Size (MB)	MSL	Dim.
all-mpnet-base-v2	109M	420	384	768
all-distilroberta-v1	82.1M	290	512	768
gte-base-en-v1.5	137M	510	8192	768

Table 2: Comparison of embedding models.

For Top2Vec and BERTopic, we experimented with three BERT-based embedding models, as shown in Table 2, to evaluate their performance. MPNet and DistilBERT require chunking to process long sequences due to their maximum sequence lengths (MSL) of 384 and 512 tokens, respectively. To better understand the impact of preprocessing and chunking strategies on topic quality, we performed a comprehensive analysis, as different strategies can significantly influence the models' effectiveness in representing long documents. Specifically, we applied mean aggregation to combine the embeddings of the text chunks, enabling the models to represent longer texts more effectively. In contrast, GTE_base can process input texts up to 8,192 tokens without chunking, making it more efficient for newspaper articles. Despite GTE_base's advantage in handling long texts, we

[2] https://spacy.io

compare all three models to assess their topic identification and coherence performance.

Implementation. We trained the classical models (LDA and NMF) on the Type 1 dataset and the neural models (Top2Vec and BERTopic) on both the Type 1 and Type 2 datasets—to identify the most suitable preprocessing strategy for neural-based TM, with a focus on overall computation time, interpretability, and the quality of the extracted topics. For LDA and NMF, we experimented with different numbers of topics, ranging from 10 to 50 in increments of 10, as these models require predefined topic counts. Although BERTopic and Top2Vec can automatically determine the number of topics, we trained these models on all three embedding models and reduced the topic count to align with LDA and NMF for a fair comparison.

Evaluation. We computed the overall computation time (in seconds) for all models, while the topics identified by each model, with varying topic counts, were processed through a separate pipeline to calculate topic coherence and diversity scores. We evaluated all the models both quantitatively and qualitatively to identify their advantages in terms of topic quality and efficiency.

4 Empirical Results & Analysis

This section outlines the experimental setup, evaluation metrics, and empirical results, followed by quantitative and qualitative analyses.

4.1 Experimental Setup

We utilized one node of a cloud-based High Performance Computing (HPC) platform to perform all of our topic-modeling experiments. The node was utilized with a configuration of 32 CPU cores, 512 GB of RAM, and one NVIDIA A100 GPU with 40 GB of VRAM. The topic model versions used are Gensim LDA (Blei et al., 2003), Gensim Online NMF (Zhao and Tan, 2017), Top2Vec version 1.0.34, and BERTopic version 0.16.3.

4.2 Evaluation Metrics

We used two different metrics: Topic Coherence and Topic Diversity. *Topic Coherence (TC)* measures the semantic similarity and logical grouping of words within a topic. The values range from -1 to 1, with higher values reflecting cohesive themes, while low scores indicate inconsistent word groupings. We utilized the Gensim Topic Coherence pipeline (Röder et al., 2015) in all our experiments. *Topic Diversity (TD)* evaluates the range of distinct topics generated by a model. We used the OCTIS Topic Diversity metric (Terragni et al., 2021b), which extracts the top-k words from each topic and aggregates the unique words across all topics. TD values range from 0 to 1, with higher scores indicating more diverse topics, while lower scores suggest redundancy among topics.

4.3 Results & Discussion

This section evaluates four topic modeling methods through both quantitative and qualitative analyses.

4.3.1 Score-Based Evaluation of Topic Models

The evaluation results summarize the performance of both classical and neural models on the Type 1 dataset and neural models on the Type 2 dataset, as shown in Tables 3 and 9. These results highlight the trade-offs between topic coherence, diversity, and computational efficiency across models and dataset configurations. When comparing classical (LDA, NMF) and neural-based (Top2Vec, BERTopic) topic-modeling approaches, it is essential to consider the inherent differences in their algorithms. Evaluating LDA and NMF separately from Top2Vec and BERTopic provides a clearer understanding of their strengths and weaknesses.

LDA and NMF show different performance patterns. NMF generally produces more coherent topics by grouping semantically similar words, while LDA excels in generating diverse topics. A notable trend with LDA is that as the number of topics increases, topic coherence decreases, thus indicating a trade-off between diversity and coherence. On the other hand, NMF maintains coherence but loses diversity with more topics, struggling to adapt to larger, more varied datasets. Both methods face challenges when the number of topics is predefined. This limitation impacts their ability to adapt to diverse datasets, demonstrating the difficulty of producing meaningful topics with fixed topic counts.

When evaluating neural-based topic models, we observed notable differences in performance across datasets and embedding models. All Top2Vec models were trained on Type 1 and 2 datasets, but none performed well across the tested embedding models. While it has the advantage of supporting various embedding models for identifying hidden themes, its overall performance was less effective than that of classical methods LDA and NMF. This suggests that, despite its flexibility, Top2Vec may

Model	#T	1955–1970			1971–1986			1987–2002			2003–2018		
		TC	*TD*	*Time*	*TC*	*TD*	*Time*	*TC*	*TD*	*Time*	*TC*	*TD*	*Time*
Classical Models													
LDA	10	**0.10**	**0.78**	**27.79**	**0.09**	**0.74**	**26.68**	**0.09**	**0.74**	**17.87**	0.04	0.68	**6.68**
	20	0.07	**0.74**	**53.10**	0.05	**0.74**	**50.70**	0.08	**0.76**	**26.70**	0.08	**0.68**	**10.42**
	30	**0.09**	**0.78**	**82.34**	0.04	**0.76**	**51.09**	0.05	**0.77**	**32.88**	0.05	**0.70**	**12.81**
	40	0.03	**0.75**	**106.05**	0.05	**0.75**	**65.91**	0.05	**0.74**	**42.62**	-0.01	**0.65**	**16.15**
	50	0.03	**0.80**	**87.89**	0.13	**0.79**	**63.79**	0.05	**0.74**	**42.62**	-0.01	**0.66**	**16.28**
NMF	10	0.08	0.77	93.80	0.08	0.71	92.59	0.08	0.81	49.59	**0.08**	**0.76**	20.78
	20	**0.08**	0.65	198.03	**0.09**	0.68	202.68	**0.10**	0.73	998.24	**0.08**	0.60	37.28
	30	0.08	0.56	229.83	**0.09**	0.60	286.73	**0.11**	0.65	114.92	**0.10**	0.52	53.83
	40	**0.09**	0.53	599.10	**0.09**	0.55	380.56	**0.12**	0.62	157.97	**0.11**	0.54	70.07
	50	**0.08**	0.49	685.23	**0.09**	0.54	486.18	**0.10**	0.55	234.19	**0.12**	0.47	113.99
Neural-based Models													
Top2Vec mpnet	10	-0.11	0.63	523.78	-0.16	0.68	450.70	-0.19	0.72	455.52	-0.14	0.74	356.75
	20	-0.12	0.52	468.59	-0.15	0.63	390.68	-0.16	0.63	380.98	-0.12	0.66	761.07
	30	-0.13	0.46	466.44	-0.12	0.56	410.65	-0.13	0.50	364.11	-0.10	0.59	161.78
	40	-0.10	0.44	467.49	-0.13	0.46	417.59	-0.12	0.51	378.59	-0.10	0.54	163.79
	50	-0.11	0.42	461.60	-0.12	0.46	413.96	-0.11	0.46	363.67	-0.11	0.50	174.21
Top2Vec distilbert	10	-0.12	0.83	340.67	-0.18	0.69	323.15	-0.20	0.69	282.86	-0.23	0.70	124.14
	20	-0.12	0.82	358.18	-0.14	0.75	295.77	-0.17	0.60	329.23	-0.18	0.53	115.40
	30	-0.12	0.83	352.39	-0.14	0.51	318.84	-0.17	0.54	333.42	-0.17	0.45	122.81
	40	-0.12	0.83	330.57	-0.15	0.48	318.19	-0.16	0.50	378.87	-0.16	0.41	127.30
	50	-0.12	0.84	359.85	-0.13	0.73	294.22	-0.15	0.48	376.95	-0.14	0.43	135.13
Top2Vec gte-base-en	10	-0.07	0.65	586.34	-0.11	0.71	630.42	-0.13	0.72	416.98	-0.09	0.66	181.92
	20	-0.06	0.53	562.34	-0.10	0.57	597.18	-0.09	0.66	472.72	-0.08	0.64	145.49
	30	-0.04	0.48	551.74	-0.08	0.51	606.02	-0.09	0.53	466.99	-0.08	0.58	144.45
	40	-0.05	0.49	562.73	-0.09	0.45	593.32	-0.08	0.48	464.97	-0.08	0.58	134.38
	50	-0.07	0.45	567.09	-0.09	0.47	577.00	-0.08	0.47	463.14	-0.08	0.52	143.61
BERTopic mpnet	10	0.16	0.83	280.39	0.07	**0.83**	218.90	**0.17**	0.90	**63.53**	**0.16**	0.88	31.82
	20	0.15	0.83	230.50	**0.14**	**0.85**	209.29	**0.14**	**0.83**	58.27	0.15	**0.83**	37.24
	30	0.15	0.76	204.14	**0.13**	**0.81**	182.67	**0.17**	0.83	60.99	0.16	0.78	34.01
	40	0.15	0.73	186.15	**0.14**	**0.77**	215.75	**0.18**	**0.79**	59.51	0.17	0.73	32.43
	50	0.15	0.70	239.87	**0.15**	0.78	**197.42**	**0.17**	0.80	62.79	0.16	0.71	35.69
BERTopic distilbert	10	**0.22**	0.83	212.30	0.08	0.78	**199.43**	0.14	0.87	254.07	0.15	**0.90**	32.49
	20	**0.21**	0.81	168.82	0.11	0.77	**192.90**	0.13	0.78	115.22	0.13	0.75	33.08
	30	**0.20**	0.77	175.23	**0.13**	0.75	267.12	0.14	0.78	200.18	0.15	0.74	36.35
	40	**0.20**	0.75	164.90	**0.15**	0.75	208.26	0.14	0.75	85.90	0.15	0.69	34.74
	50	**0.22**	0.74	181.44	**0.15**	0.71	261.99	0.16	0.76	62.89	0.14	0.69	34.90
BERTopic gte-base-en	10	0.15	**0.86**	**110.89**	**0.12**	**0.90**	255.98	0.15	**0.92**	80.00	0.15	0.88	**28.97**
	20	0.14	**0.88**	**79.56**	**0.14**	0.83	291.60	0.13	0.82	**49.31**	**0.16**	0.79	**31.30**
	30	0.15	**0.84**	**85.04**	**0.13**	0.78	**176.11**	0.14	**0.85**	**52.29**	**0.18**	**0.79**	**30.63**
	40	0.16	**0.79**	**84.35**	**0.14**	**0.77**	**168.87**	0.14	**0.79**	**56.03**	**0.18**	**0.77**	**26.76**
	50	0.16	**0.77**	**81.82**	**0.15**	**0.79**	235.22	0.16	**0.81**	**52.40**	**0.18**	**0.75**	**31.15**

Table 3: Quantitative Results for the LDA, NMF, Top2Vec, BERTopic Performance Scores on the Type 1 Dataset across different numbers of topics (#T)

not be the optimal choice for large-scale datasets due to inefficiencies in both topic quality and diversity. When analyzing the results of BERTopic models trained on both Type 1 and 2 datasets, we found that models trained on Type 1 data outperformed all other models, as shown in Table 3. BERTopic consistently achieved higher TC and TD scores on Type 1 compared to Type 2 (Table 9), where it performed less effectively, especially with fewer topics and performance improved with more topics. This poor performance is likely due to the presence of stop words that affect the topic formation. This underscores the importance of post-processing techniques to refine results.

Findings. All three BERTopic variants trained on Type 1 data outperformed LDA, NMF, and Top2Vec, with stable performance across different topic ranges, highlighting the crucial role of preprocessing and embedding models in generating high-quality contextual representations. Specifically, Type 1 preprocessing—which included text normalization, stopword removal, and lemmatization—enhanced topic coherence by reducing noise and improving semantic consistency, while minimal preprocessing resulted in noisier topic distributions and lower coherence scores.

These findings underscore the importance of selecting preprocessing strategies suited to the dataset

No.	Topic Words
1	people, world, country, war, time, man, long, know, mean, problem
2	radiation, radioactive, health, doctor, radioactivity, medical, disease, use, effect, patient
3	united, nuclear, states, weapon, disarmament, conference, soviet, american, agreement, president
4	military, weapon, defense, army, rocket, force, air, missile, equip, aircraft
5	european, common, market, europe, economic, country, community, trade, brussels, euratom
6	council, vote, session, committee, assembly, member, president, commission, international, general
7	franc, tax, council, million, construction, increase, state, federal, canton, new
8	man, know, day, church, english, like, time, come, want, war
9	use, device, meter, water, high, time, machine, gas, light, temperature
10	time, year, water, use, work, know, new, waste, long, life

Table 4: List of 10 topics out of 50 discovered by LDA.

No.	Topic Words
1	plant, company, construction, swiss, power, water, zurich, electricity, industry, switzerland
2	france, french, gaulle, general, europe, paris, european, force, political, nuclear
3	economic, industry, economy, trade, market, company, development, policy, sector, berlin
4	states, united, nuclear, test, american, ussr, experiment, explosion, agreement, washington
5	reactor, research, atomic, uranium, new, water, scientific, project, center, carry
6	council, vote, session, committee, assembly, member, president, commission, international, general
7	million, increase, company, year, price, share, franc, billion, production, bank
8	car, accident, fire, police, year, injure, zurich, die, road, people
9	work, school, study, university, institute, technical, research, professor, use, service
10	water, war, man, want, long, new, peace, west, like, come

Table 5: List of 10 topics out of 50 discovered by NMF.

and the assumptions of different topic models. Neural models like BERTopic benefit from structured preprocessing, which refines input representations and improves topic extraction. To optimize topic modeling performance, we recommend either structured preprocessing for neural models or minimal preprocessing combined with robust post-processing techniques like topic merging and filtering.

4.3.2 Computational Efficiency & Scalability

We focus only on Table 3, as models trained on Type 2 with minimal preprocessing exhibited poor performance. The computational demands of each model vary depending on their underlying algorithms. The classical models (LDA and NMF) rely on probabilistic inference and matrix factorization, respectively, thereby requiring multiple iterative updates. As the number of topics increases, their computational cost grows significantly, leading to longer training times. In contrast, neural-based models like Top2Vec and BERTopic use pre-trained embeddings and clustering techniques, allowing automatic determination of the optimal number of topics, and improving scalability without manual intervention. However, our experiments revealed that Top2Vec exhibited a significantly higher computational cost than classical methods across all tested embedding models. Despite its flexibility in supporting different SBERT variants, it proved to be computationally expensive and less scalable for very large datasets. On the other hand, BERTopic demonstrated superior computational efficiency, leveraging transformer-based embeddings and clustering techniques to extract high-quality topics with stable computation time. This efficiency, combined with strong performance, makes BERTopic a scalable and reliable choice for large datasets, particularly with appropriate preprocessing.

Findings. Overall, selecting the right TM approach requires balancing performance and computational efficiency. Our experiments suggest that BERTopic, with its strong topic coherence, diversity, and manageable computational demands, is the preferred choice for scalable and high-quality TM.

4.3.3 Topic Interpretability & Quality

Our numerical results show that LDA excelled in topic diversity, while NMF performed better in topic coherence. However, BERTopic outperformed by generating more coherent and diverse topics simultaneously. Additionally, we qualitatively analyzed these models that performed well in numerical evaluations, now focusing on the quality and relevance of the generated topics.

Tables 4 and 5 show the topics identified by

No.	Topic Words
1	nuclear, weapon, disarmament, conference, soviet, united, treaty, states, atomic, agreement
2	church, pope, world, god, catholic, man, bishop, cardinal, people, peace
3	energy, plant, reactor, power, atomic, nuclear, electricity, use, construction, uranium
4	council, federal, music, swiss, franc, year, national, million, new, work
5	chinese, china, beijing, communist, mao, soviet, nuclear, moscow, bomb, party
6	crash, plane, accident, aircraft, pilot, air, bomb, meter, near, flight
7	india, nehru, indian, chinese, delhi, china, border, minister, prime, new
8	diefenbaker, canadian, canada, pearson, party, liberal, ottawa, government, lester, quebec
9	japanese, japan, okinawa, sato, tokyo, asia, states, american, united, kishi
10	car, accident, fire, police, year, injure, zurich, die, road, people

Table 6: List of 10 topics out of 50 discovered by BERTopic-MPNET.

No.	Topic Words
1	nuclear, united, new, soviet, government, country, year, states, american, state
2	church, man, world, life, people, human, work, time, god, war
3	energy, reactor, plant, power, atomic, nuclear, use, electricity, construction, research
4	radiation, radioactive, radioactivity, atomic, danger, effect, explosion, nuclear, waste, bomb
5	chinese, china, beijing, communist, mao, nuclear, soviet, moscow, bomb, party
6	crash, plane, aircraft, pilot, accident, air, bomb, meter, near, flight
7	india, nehru, indian, minister, china, pakistan, shastri, delhi, prime, nuclear
8	canadian, diefenbaker, canada, pearson, party, liberal, government, ottawa, election, quebec
9	japanese, japan, okinawa, tokyo, nuclear, hiroshima, sato, american, united, states
10	conference, session, stop, testing, nuclear, weapon, draft, article, delegate, delegation

Table 7: List of 10 topics out of 50 discovered by BERTopic-DistilBERT.

No.	Topic Words
1	disarmament, conference, soviet, nuclear, united, agreement, treaty, states, weapon, geneva
2	church, peace, pope, world, people, man, war, council, easter, bishop
3	energy, plant, reactor, power, atomic, electricity, nuclear, construction, switzerland, swiss
4	radioactive, radiation, radioactivity, use, atomic, effect, bomb, cancer, human, danger
5	explosion, bomb, chinese, test, nuclear, china, atomic, experiment, carry, french
6	crash, plane, bomb, aircraft, pilot, accident, air, bomber, b52, flight
7	spy, espionage, frauenknecht, agent, secret, affair, soviet, trial, service, penkovsky
8	council, federal, franc, swiss, year, canton, zurich, national, vote, councilor
9	japanese, japan, okinawa, sato, tokyo, china, kishi, asia, american, island
10	french, strike, force, france, government, national, paris, minister, gaullist, pompidou

Table 8: List of 10 topics out of 50 discovered by BERTopic-GTE_base.

LDA and NMF, respectively. LDA performs better in computation time but generates more generic topics that lack meaningfulness, particularly the last three topics highlighted in "*red*". This is due to LDA's fixed number of topics, which does not adapt well to large, heterogeneous datasets, leading to reduced topic quality. In contrast, NMF requires more computation time but produces more logical and coherent topics, though some generic topics, like Topic_10 highlighted in "*red*", still appear. Exploring all 50 topics from NMF reveals redundancy, likely caused by the fixed topic count, which limits adaptation to the data. This suggests that while NMF excels in quality, it may suffer from overfitting or redundancy with too many topics. We recommend NMF over LDA for more meaningful topics, especially with fewer topics. However, a high topic count may lead to redundancy, so balancing topic number and performance is crucial.

The sample list of 10 topics produced by the MPNET, DistilBERT, and GTE_base variants of the BERTopic models is shown in Tables 6, 7, and 8 with distinct topics (in *black*) and most similar topics highlighted using different colors. Comparing Tables 6 and 7, we observe only few topics are distinct, and most are similar topics, with slight variations in their word compositions. This indicates that the embeddings generated by both models are quite similar, leading to overlapping topic generation. In contrast, GTE_base (Table 8) generates topics that blend words from both MPNET and DistilBERT, but with better and more

meaningful topic representations. For instance, in Topic_2, GTE_base identifies the words "easter" and "council", which are missing in both MPNET and DistilBERT. This demonstrates GTE_base's ability to capture more specific and contextually relevant terms, such as those related to a council associated with Easter, resulting in a more coherent interpretation of the topic. In contrast, MPNET and DistilBERT miss this connection, suggesting GTE_base's advantage in understanding subtle contextual relationships within the text. Similarly, Topic_6 from GTE_base captures the keywords "B52" and "bomber", which refer to the American long-range strategic bomber. These terms are not present in the other two models, further showcasing GTE_base's capacity to capture specific, contextually rich terms that may be crucial for understanding the historical context of the topics.

Findings. Although the quantitative results for neural models are similar, they do not capture nuanced differences in topic relevance and coherence, emphasizing the need for qualitative analysis. While MPNET and DistilBERT can be improved with advanced chunking and aggregation strategies, GTE_base's ability to handle longer sequences makes it better suited for topic modeling tasks, especially when dealing with long texts.

5 Conclusions

In this paper, we conducted a comprehensive evaluation of four topic-modeling techniques—LDA, NMF, Top2Vec, and BERTopic—in combination with three text-embedding models. While our experiments leverage HPC for large datasets, all tested methods remain effective on standard hardware for smaller datasets, ensuring accessibility and scalability across diverse computational settings. Our experiments show that LDA excels in topic diversity but struggles with coherence, while NMF generates more coherent topics but suffers from redundancy with a large number of topics. BERTopic with a large sequence-length embedding model outperforms both, offering superior coherence, diversity, and the ability to handle longer texts without losing context. We recommend BERTopic for large, heterogeneous datasets due to its balance of efficiency, coherence, and diversity, although careful preprocessing is necessary for models like smaller embeddings models. Our empirical analysis provides clear guidance for digital humanities researchers and users in selecting the most appropriate topic modeling method for their specific use cases, particularly when dealing with large datasets.

Limitations

Our current work is limited to the original LDA and NMF variants, and the performance of other variants remains to be tested. In future work, we plan to explore BERTopic with recent LLM-based embeddings to enhance topic representation and improve clustering accuracy, as well as investigate other BERT-based models with alternative chunking strategies. Additionally, we aim to incorporate dynamic topic modeling to capture the evolution of topics over time, enabling a more nuanced understanding of temporal trends. We have already conducted preliminary experiments in this direction and intend to further refine and evaluate the approach.

References

Robert B. Allen and Robert Sieczkiewicz. 2010. How historians use historical newspapers. In *Proceedings of the 73rd Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, USA. American Society for Information Science.

Angela Ambrosino, Mario Cedrini, John B Davis, Stefano Fiori, Marco Guerzoni, and Massimiliano Nuccio. 2018. What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology*, 25(4):329–348.

Dimitar Angelov. 2020. Top2vec: Distributed representations of topics. *ArXiv*, abs/2008.09470.

Alina Arseniev-Koehler, Susan D. Cochran, Vickie M. Mays, Kai Wei Chang, and Jacob Gates Foster. 2020. Integrating topic modeling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences of the United States of America*, 119.

Anda Baklāne and Valdis Saulespurēns. 2022. The application of latent dirichlet allocation for the analysis of latvian historical newspapers: Oskars kalpaks' case study. *Science, technologies, innovation*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Svetlana S. Bodrunova. 2021. *Topic Modeling in Russia: Current Approaches and Issues in Methodology*, pages 409–426. Springer International Publishing, Cham.

Nicolas Bourgeois, Aurélien Pellet, and Marie Puren. 2022. Using topic generation model to explore the

french parliamentary debates during the early third republic (1881-1899). In *Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop co-located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, volume 3133, pages 35–51.

Marc Chappelle, Sakaria Laisene Auelua-Toomey, and Steven O. Roberts. 2024. Sankofa: Using topic models to review the history of the journal of black psychology. *Journal of Black Psychology*, 50(1):9–29.

Kostadin Cvejoski, Ramsés J. Sánchez, and C. Ojeda. 2023. Neural dynamic focused topic model. In *AAAI Conference on Artificial Intelligence*.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41:391–407.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7.

Mats Fridlund and René Brauer. 2013. Historizing topic models: A distant reading of topic modeling texts within historical studies. In *Cultural Research in the Context of Digital Humanities*, pages 152–63. Herzen State Pedagogical University.

Michael Gavin and Eric Gidal. 2016. Topic modeling and the historical geography of scotland. *Studies in Scottish Literature*, 42(2):185–197.

Michael Ginn and Mans Hulden. 2024. Historia magistra vitae: Dynamic topic modeling of roman literature using neural embeddings. *arXiv preprint arXiv:2406.18907*.

Philip Grant, Ratan Sebastian, Marc Allassonnière-Tang, and Sara Cosemans. 2021. Topic modelling on archive documents from the 1970s: global policies on refugees. *Digital Scholarship in the Humanities*, 36(4):886–904.

Maria Gabriella Grassia, Marina Marino, Rocco Mazza, Michelangelo Misuraca, Agostino Stavolo, et al. 2022. Topic modeling for analysing the russian propaganda in the conflict with ukraine. *ASA 2022*, page 245.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Ekaterina Gryaznova and Margarita Kirina. 2021. Defining kinds of violence in russian short stories of 1900-1930: A case of topic modelling with lda and pca. In *IMS*, pages 281–290.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii. Association for Computational Linguistics.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Conference on Uncertainty in Artificial Intelligence*.

Eirini Karamouzi, Maria Pontiki, and Yannis Krasonikolakis. 2024. Historical portrayal of Greek tourism through topic modeling on international newspapers. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 121–132, St. Julians, Malta. Association for Computational Linguistics.

Salima Lamsiyah, Keerthana Murugaraj, and Christoph Schommer. 2023. Historical-domain pre-trained language model for historical extractive text summarization.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

King Ip Lin and Sabrina Peng. 2022. Enhancing digital history – event discovery via topic modeling and change detection. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 69–78, Taipei, Taiwan. Association for Computational Linguistics.

Anna Maltseva, Natalia Shilkina, Evgeniy Evseev, Mikhail Matveev, and Olesia Makhnytkina. 2021. Topic modeling of russian-language texts using the parts-of-speech composition of topics (on the example of volunteer movement semantics in social media). In *2021 29th Conference of Open Innovations Association (FRUCT)*, pages 247–253.

Jani Marjanen, Elaine Zosa, Simon Hengchen, Lidia Pivovarova, and Mikko Tolonen. 2020. Topic modelling discourse dynamics in historical newspapers. *ArXiv*, abs/2011.10428.

Ginevra Martinelli, Paola Impicciché, Elisabetta Fersini, Francesco Mambrini, and Marco Passarotti. 2024. Exploring neural topic modeling on a classical Latin corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6929–6934, Torino, Italia. ELRA and ICCL.

Keerthana Murugaraj, Salima Lamsiyah, and Christoph Schommer. 2025. Abstractive summarization of historical documents: A new dataset and novel method using a domain-specific pretrained model. *IEEE Access*, 13:10918–10932.

Mila Oiva. 2020. Topic modeling russian history. *The Palgrave Handbook of Digital Russia Studies*.

Martin Orr, Kirsten Van Kessel, and David Parry. 2024. Ethical thematic and topic modelling analysis of sleep concerns in a social media derived suicidality dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 74–91, St. Julians, Malta. Association for Computational Linguistics.

Swati Rajwal, Avinash Kumar Pandey, Zhishuo Han, and Abeed Sarker. 2024. Unveiling voices: Identification of concerns in a social media breast cancer cohort via natural language processing. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 264–270, Torino, Italia. ELRA and ICCL.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021b. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Ana Sabina Uban, Cornelia Caragea, and Liviu P. Dinu. 2021. Studying the evolution of scientific topics and their relationships. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1908–1922, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Tze-I Yang, Andrew Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.

Ekaterina Zamiraylova and Olga Mitrofanova. 2020. Dynamic topic modeling of russian prose of the first third of the xxth century by means of non-negative matrix factorization. In *Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings*, volume 2552, pages 321–339.

Renbo Zhao and Vincent Y. F. Tan. 2017. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing*, 65(3):555–570.

APPENDIX

Model	#T	1955–1970			1971–1986			1987–2002			2003–2018		
		TC	*TD*	*Time*	*TC*	*TD*	*Time*	*TC*	*TD*	*Time*	*TC*	*TD*	*Time*
Type 2: Minimally Preprocessed Dataset													
Top2Vec mpnet	10	-0.15	0.58	786.86	-0.16	0.72	679.30	-0.14	0.71	736.28	-0.17	0.70	221.45
	20	-0.12	0.57	756.84	-0.16	0.57	687.96	-0.14	0.57	758.25	-0.12	0.61	246.58
	30	-0.15	0.51	749.55	-0.14	0.53	680.66	-0.14	0.56	743.47	-0.11	0.57	263.75
	40	-0.12	0.51	794.00	-0.13	0.52	677.60	-0.13	0.53	613.16	-0.11	0.55	242.13
	50	-0.12	0.49	760.83	-0.14	0.49	700.08	-0.13	0.53	574.69	-0.11	0.53	274.99
BERTopic distilbert	10	0.002	0.42	350.03	-0.03	0.30	148.19	0.005	0.378	98.50	0.029	0.41	35.87
	20	0.001	0.32	320.78	-0.02	0.29	121.18	0.04	0.32	72.40	0.03	0.35	37.72
	30	0.001	0.31	372.43	-0.0002	0.31	127.89	0.022	0.40	162.45	0.04	0.38	35.33
	40	0.01	0.38	376.79	0.015	0.39	122.81	0.020	0.36	71.17	0.06	0.43	36.43
	50	0.03	0.39	381.09	0.033	0.43	127.32	0.04	0.47	98.58	0.06	0.45	39.09
Top2Vec gte-base-en	10	-0.13	0.62	839.35	-0.13	0.77	830.23	-0.08	0.75	488.67	-0.10	0.69	198.65
	20	-0.09	0.59	839.45	-0.10	0.63	853.75	-0.11	0.63	505.56	-0.10	0.62	192.38
	30	-0.09	0.55	827.68	-0.10	0.57	836.90	-0.12	0.58	526.19	-0.12	0.56	194.46
	40	-0.08	0.56	837.36	-0.09	0.57	845.31	-0.11	0.49	475.23	-0.11	0.52	198.48
	50	-0.08	0.48	830.75	-0.08	0.54	832.88	-0.10	0.46	475.27	-0.12	0.54	196.63
BERTopic mpnet	10	-0.01	0.34	475.08	-0.02	0.29	106.50	-0.003	0.37	132.66	0.034	0.47	71.75
	20	-0.001	0.27	389.85	-0.013	0.29	119.57	0.006	0.32	72.23	0.024	0.34	39.78
	30	-0.002	0.31	403.05	0.008	0.36	116.64	0.007	0.34	73.98	0.04	0.39	36.98
	40	0.003	0.33	401.80	0.02	0.39	124.75	0.02	0.40	72.96	0.05	0.41	39.17
	50	0.022	0.38	333.86	0.024	0.40	119.11	0.029	0.41	71.90	0.06	0.43	36.25
BERTopic distilbert	10	0.002	0.42	350.03	-0.03	0.30	148.19	0.005	0.378	98.50	0.029	0.41	35.87
	20	0.001	0.32	320.78	-0.02	0.29	121.18	0.04	0.32	72.40	0.03	0.35	37.72
	30	0.001	0.31	372.43	-0.0002	0.31	127.89	0.022	0.40	162.45	0.04	0.38	35.33
	40	0.01	0.38	376.79	0.015	0.39	122.81	0.020	0.36	71.17	0.06	0.43	36.43
	50	0.03	0.39	381.09	0.033	0.43	127.32	0.04	0.47	98.58	0.06	0.45	39.09
BERTopic gte-base-en	10	0.03	0.57	170.98	-0.02	0.33	137.64	-0.002	0.37	103.49	0.02	0.44	60.91
	20	0.03	0.45	120.92	0.01	0.41	148.14	0.008	0.34	64.04	0.05	0.41	32.72
	30	0.03	0.46	119.42	0.02	0.39	291.37	0.03	0.43	65.52	0.07	0.44	34.46
	40	0.03	0.46	122.54	0.03	0.42	371.86	0.030	0.40	62.95	0.06	0.46	34.48
	50	0.04	0.46	122.64	0.04	0.47	234.56	0.05	0.47	66.20	0.07	0.48	34.83

Table 9: Quantitative Results for the Neural Topic Models (Top2Vec and BERTopic) on the Type-2 Dataset.