Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 4, August 2025, Page. 1957-1970
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5090

Enhancing BERTopic with Neural Network Clustering for Thematic Analysis of U.S. Presidential Speeches

Sajarwo Anggai*[1], Rafi Mahmud Zain[2], Tukiyat[3], A. A Waskita [4]

[1,2,3]Graduate Program of Informatics Engineering, Universitas Pamulang, Indonesia
[3,4]Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia

Email: [1]sajarwo@gmail.com, [2]rafizain777@gmail.com, [3]dosen02711@unpam.ac.id, [4]arya.adhyaksa.waskita@brin.go.id

Received: Jul 13, 2025; Revised: Aug 12, 2025; Accepted: Aug 13, 2025; Published: Aug 18, 2025

Abstract

Understanding the underlying themes in presidential speeches is critical for analyzing political discourse and determining public policy direction. However, topic modeling in this context presents difficulties, particularly when clustering semantically rich topics from high-dimensional embeddings. This study seeks to improve topic modeling performance by incorporating a Neural Network Clustering (NNC) approach into the BERTopic pipeline. We analyze 2,747 speeches delivered by U.S President Joe Biden (2021-2025) and compare three clustering techniques: HDBSCAN, KMeans, and the proposed Autoencoder-based NNC. The evaluation metrics (UMass, NPMI, Topic Diversity) show that NNC produces the most coherent and diverse topic clusters (UMass = -0.4548, NPMI = 0.0234, Diversity = 0.3950,). These findings show that NNC can overcome the limitations of density and centroid-based clustering in high-dimensional semantic spaces. The study contributes to the field of Natural Language Processing by demonstrating how neural-based clustering can improve topic modeling, particularly for complex, real-world political corpora.

Keywords : *Autoencoder, BERTopic, Deep Clustering, Political Discourse, Topic Modeling.*

1. INTRODUCTION

Advancements in information technology have significantly transformed human access to information and data processing, particularly within the realm of political and state communication. The official speech of the head of state stands as a significant product of political communication. It serves as a prominent and transparent medium, functioning not merely as a political vehicle but also reflecting the nation's political orientation, governance style, and socio-economic attributes. Presidential speeches serve as strategic texts, embedding both implicit and explicit messages, thus warranting systematic analysis from various perspectives[1]. In the realm of text-based investigation, Topic Modeling has emerged as a widely utilized method for identifying the primary themes within document collections, all without the need for explicit labeling. A contemporary method that delivers exceptional results in topic extraction is BERTopic, which integrates Transformer-based embeddings with dimensionality reduction and clustering strategies[2]. BERTopic provides a significantly improved semantic representation of documents when compared to conventional techniques like Latent Dirichlet Allocation[3]. However, the predominant clustering method utilized in BERTopic, HDBSCAN, falls short in its capacity to uncover more intricate latent structures, particularly within high-dimensional datasets[4].

To address these limitations, this study integrates BERTopic with an Autoencoder-based Neural Network Clustering (NNC) approach. This method leverages the capacity of artificial neural networks to capture non-linear relationships among documents, transforming them into a lower-dimensional latent space, and subsequently applies the KMeans algorithm for clustering[5]. Consequently, the process of identifying topics is refined and organized effectively.

This study analyzes the official speech of the President of the United States, selected for its historically significant global policy content. The information utilized is derived from the transcripts of President Joe Biden's speeches throughout his term from 2021 to 2025.

This study employs a combination of BERTopic and NNC to uncover the primary themes present in the speeches. Additionally, it seeks to compare the modeling outcomes generated by traditional methods like HDBSCAN with those produced by the proposed Neural Network Clustering approach.

This approach aims to enhance the development of techniques for analyzing policy texts through topic modeling and deep clustering, thereby reinforcing a data-driven understanding of official political discourse and the formulation of government policies.

2. METHOD

This study presents an integrated approach that combines neural network–based clustering, specifically Deep Embedded Clustering (DEC), with the BERTopic framework to analyze thematic patterns in U.S. presidential speeches. The process begins with transcript preprocessing, followed by the generation of semantic sentence embeddings using a transformer-based language model. These embeddings are clustered into semantically coherent groups using DEC, which are then processed through BERTopic to extract and visualize key themes using class-based TF-IDF (c-TF-IDF). This framework enables the identification of evolving discourse trends across election cycles, offering insights into shifting public sentiment and political priorities. The model’s performance is evaluated through topic coherence and diversity metrics, demonstrating its effectiveness in enhancing the interpretability and accuracy of thematic analysis in political speech.

The overall architecture of this research workflow is illustrated in Figure 1. It outlines the sequential steps starting from data collection and preprocessing, followed by embedding and dimensionality reduction. Three clustering methods HDBSCAN, KMeans, and NNC are applied to the reduced embeddings. The resulting topic structures are evaluated to compare performance across clustering strategies, particularly in terms of coherence and topic diversity.

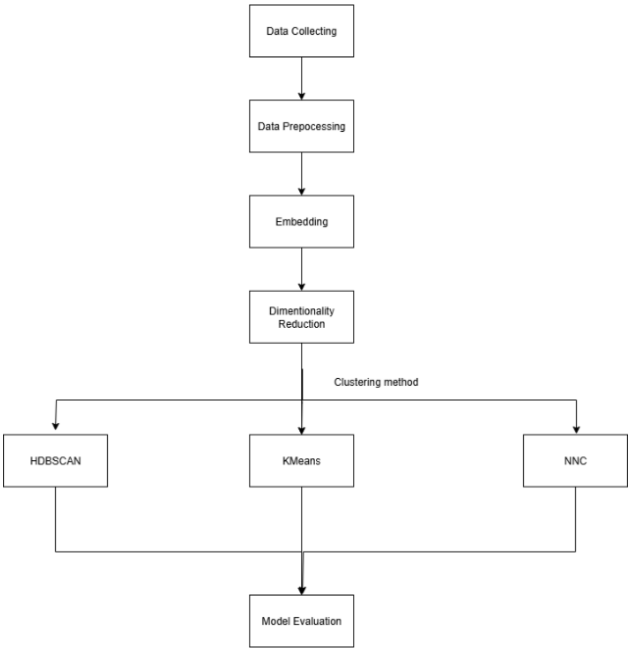

Figure 1. Flowchart Research

2.1. Data Collection Technique

The data was obtained through the application of web scraping methods[6]. This method extracts the complete text from the specified website by modifying the HTML structure and systematically storing it in a MySQL database.

The extraction process was conducted utilizing Python alongside the BeautifulSoup4 library, enabling the automated retrieval of text from a website according to its HTML framework. The collected data is preserved in SQL format for subsequent analysis and processing.

2.2. Pre-Processing

Prior to Topic Modeling, the data needs to undergo multiple preprocessing stages. The purpose of this preprocessing is to organize the data and guarantee that only pertinent words are examined[7]. Preprocessing plays a crucial role in converting text from human language into a format that machines can understand in text mining methodologies [8]. The preprocessing stage is crucial for organizing unstructured text and retaining key terms that aid in defining the topic category of the text[1][9]. Natural language text often includes numerous words that lack specific meaning, including prepositions and pronouns. To streamline the text data, purify the data, and minimize interference.

2.3. Topic Modeling

Topic modeling is an unsupervised machine learning method that facilitates the automatic identification of abstract topics or themes within an extensive corpus of documents[10][11]. It is extensively utilized in natural language processing (NLP) and text mining to discern latent semantic structures present in a corpus, without requiring pre-labeled data. Topic modeling operates under the premise that documents comprise a combination of latent topics, each characterized by a distribution of words. Conventional models like LDA examine word co-occurrence to deduce structure, whereas contemporary methods such as BERTopic improve upon this by integrating contextual embeddings to produce more coherent and significant topics[12].

2.4. BERTopic

BERTopic is a model based on Transformers that comprehensively grasps text context in both directions, leading to text representations that are deeply infused with semantic significance[13] [14]. This is employed in BERTopic to construct topic representations via a sequence of methodical steps. This framework integrates word embedding, dimension reduction, and clustering algorithms to create semantically similar clusters[15]. This method utilizes BERT's capabilities to generate superior document embeddings, facilitating more precise and pertinent topic identification.

2.5. Transformers Embedding

BERTopic begins by converting our input papers into numerical forms[16] [17]. Though there are several methods to do this, we usually employ sentence-transformers ("all-MiniLM-L6-v2") since they are rather good at identifying the semantic similarity between papers.

2.6. Dimentionality Reduction

The document or textual data that has been transformed into a vector will undergo the clustering process. The initial step involves dimensionality reduction, commonly referred to as dimension reduction, utilizing the Uniform Manifold Approximation and Projection (UMAP)[18]. Dimensionality reduction involves a technique aimed at minimizing the volume of data that persists following the creation of embeddings. This process will produce data that is akin to a vector, albeit with reduced dimensionality.

2.7. Clustering

This study employs an autoencoder-based NNC approach as a substitute for traditional clustering techniques like HDBSCAN[4][5]. An autoencoder comprises a deep neural network architecture featuring two primary components: the encoder and the decoder[19]. The encoder transforms a high-dimensional input into a lower-dimensional latent representation, whereas the decoder strives to reconstruct the original input from that representation[20][21][22].

The encoder function can be expressed mathematically by Equation (1):

$$z = f(x) = \sigma(W_{\mathrm{e}} \cdot x + b_{\mathrm{e}}) \quad (1)$$

where:

- $x \in \mathbb{R}^{\mathrm{n}}$ = input feature vector from document embeddings
- $z \in \mathbb{R}^{\mathrm{m}}$ = lower dimensional latent representation with m < n
- W_{e} = weighted encoder
- b_{e} = bias *encoder*
- σ = non-linear activation function

The decoder is in charge of reconstructing the original data from the latent representation. The decoder function is given by Equation (2):

$$\hat{x} = g(z) = \sigma(W\_d \cdot z + b\_d) \quad (2)$$

The Autoencoder training process aims to minimise the loss of the reconstruction function, generally using the mean squared error. This loss function is defined in Equation (3):

$$L\_rec(x, \hat{x}) = \sum_{\mathrm{i}} (x_{\mathrm{i}} - \hat{x}_{\mathrm{i}})^2 \quad (3)$$

Upon completion of the training process and achieving model stability, the output from the latent layer serves as a novel representation for the documents. The representation is subsequently clustered utilizing the KMeans algorithm, focusing on minimizing the distance between documents and cluster centers. The objective function of KMeans is expressed in Equation (4):

$$arg\ min\_C \sum_{\mathrm{i}} = 1\hat{}k \sum\_\{x_{\mathrm{j}} \in C_{\mathrm{i}}\} ||x_{\mathrm{j}} - \mu_{\mathrm{i}}||^2 \quad (4)$$

This method involves processing the embedding result vector generated by the model transformer (BERTopic) through an Autoencoder, which uncovers the latent structure of the data, thereby enhancing its suitability for clustering. Subsequently, a KMeans algorithm is employed to categorize the representation into more distinct and meaningful topic clusters. In contrast to density-based clustering methods like HDBSCAN, which are influenced by parameters such as minimum cluster size and distance threshold, the NNC approach effectively captures the non-linear relationships between documents, enhancing cluster separability in low-dimensional feature spaces.

2.8. Model Evaluation

In assessing the quality of topics produced by different topic modeling approaches, the author employs two coherence score metrics: UMASS and NPMI. The two metrics commonly serve to evaluate the semantic coherence among words within a given topic[8].

2.8.1 UMass

The UMas metric assesses the coherence score by analyzing the frequency of word occurrences within the same topic throughout the document[23]. UMass assesses the likelihood of a word pair co-occurring within a document and contrasts it with the individual likelihood of one of the words in the pair[24].

The mathematical representation of UMass is calculated using Equation (5):

$$C_{UMass}(T) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \log \frac{p(w_i,w_j) + \frac{1}{D}}{p(w_j)} \qquad (5)$$

where:

- T={w1,w2,…,wN} : list of words in topic.
- p(wi,wj) : the probability of two words wi and wj co-occurring in the document (context).
- p(wj) : the probability of word wj appearing in the document.
- N : word count in the topic.
- D : total number of documents in the corpus.

2.8.2 NPMI

In addition to UMASS, we utilize Normalised Pointwise Mutual Information (NPMI)[25]. This metric assesses the degree of association between words within a topic relative to their standalone occurrences[26]. NPMI computes the Pointwise Mutual Information (PMI) in a normalized manner, ensuring that the outcome falls within the range of -1 to 1[27]. NPMI is formulated as shown in Equation (6):

$$NPMI(w_i, w_j) = \left(\frac{PMI}{-\log(p(w_i,w_j) + \epsilon)} \right) \qquad (6)$$

The equation formulates the rate at which words co-occur in a given corpus, known as PMI is calculated by Equation (7):

$$PMI(w_i, w_j) = \log \frac{(p(w_i,w_j) + \epsilon)}{(p(w_i).\ p(w_j)} \qquad (7)$$

where:

- p(wi,wj) : the probability of two words wi and wj co-occurring in the corpus.
- p(wi) : the probability of the word wi appearing in the corpus.
- p(wj) : the probability of word wj appearing in the corpus.
- ϵ : small value to prevent the logarithm of zero.

3. RESULT

3.1. Data Collection

This study utilizes a corpus composed exclusively of official speeches delivered by U.S. President Joe Biden, sourced from the official White House website. A total of 2,747 speech transcripts were collected, covering the full duration of his presidential term from January 20, 2021, to January 19, 2025. These textual transcripts serve as the primary dataset for the topic modeling analysis conducted in this study

3.2. Data Prepocessing Result

Following an extensive preprocessing phase of the speeches delivered by the President, Vice President, and First Lady, a total of 2,747 cleaned textual documents were successfully extracted and prepared for topic modeling. The preprocessing pipeline included several standard natural language processing steps, such as converting all text to lowercase, removing punctuation, numeric characters, and stopwords, and applying n-gram-based tokenization to preserve multi-word expressions. These steps aimed to reduce noise while retaining semantically meaningful content. The cleaned text was then stored in a dedicated column labeled *Preprocessed_Content*, which served as the primary input for the subsequent embedding and topic modeling processes.

3.3. Topic Result

During the topic modeling process, three distinct clustering algorithms were employed alongside the BERTopic framework to uncover latent themes within the corpus: HDBSCAN, KMeans, and NNC. Each model was configured to generate 20 topics, allowing for a fair comparison across the different approaches.

3.3.1. HDBSCAN Method

Table 1. Result of HDBSCAN

Topic	Count	Name
-1	787	worker_lead_folk_create
0	291	vice_president_freedom_vice_fight
1	269	vote_deal_president_happen
2	203	security_economic_climate_continue
3	186	folk_percent_pay_big
4	150	woman_care_issue_business
5	146	economy_cost_economic_plan
6	126	school_business_investment_future
7	103	life_stand_bring_live
8	88	climate_administration_issue_investment
9	84	support_security_continue_stand
10	70	help_state_home_sure
11	65	life_family_nation_climate
12	46	percent_month_continue_school
13	41	great_history_home_thank thank
14	26	vice president_vice_business_fight
15	19	security_continue_end_issue
16	17	care_freedom_fight_vice president
17	15	child_home_family_protect
18	15	woman_care_lead_life

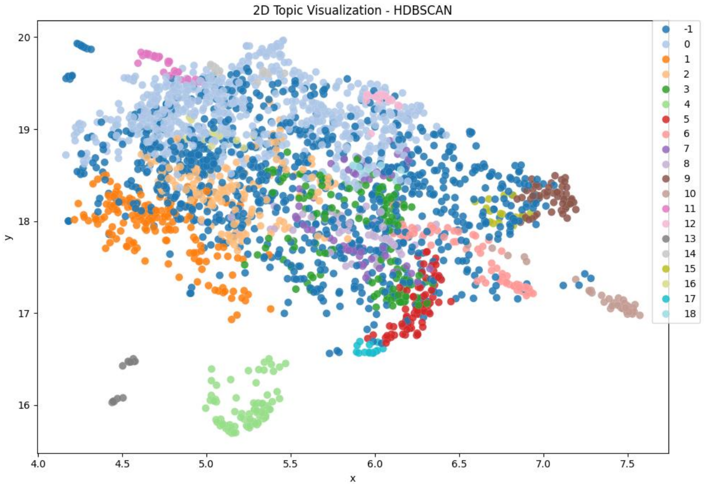

Figure 2. HDBSCAN Plot

The clustering approach using HDBSCAN produced 20 distinct topics, along with a significant number of documents (787 instances) classified as outliers under topic -1, indicating content that did not strongly align with any specific cluster. The analysis identified recurring themes such as freedom, economic policy, healthcare, and investment in education based on the topics generated by HDBSCAN. These results are summarized in Table 1, while the topic distribution and outlier separation are visualized in Figure 1.

3.3.2. KMeans Method

Meanwhile, the KMeans clustering approach produced a more balanced distribution of topic assignments, without generating any outlier labels. The topics identified through KMeans analysis similarly revealed recurring themes related to security, the economy, education, and women's issues, indicating a level of thematic consistency that extends across cluster boundaries. These results are summarized in Table 2, while the topic distribution are visualized in Figure 2.

Table 2. Result of KMeans

Topic	Count	Name
0	107	security_support_continue_stand
1	185	percent_pay_folk_lot
2	128	worker_business_investment_administration
3	164	deal_president_able_happen
4	87	plan_make sure_sure_care
5	161	life_family_live_history
6	182	freedom_care_vote_woman
7	249	freedom_life_nation_history
8	109	school_child_community_life
9	111	folk_pay_worker_big
10	185	woman_care_issue_law
11	51	percent_school_month_continue
12	44	great_home_history_use
13	145	economy_economic_cost_plan
14	179	vice president_vice_freedom_fight
15	92	climate_investment_economic_new
16	142	vote_vice president_happen_believe
17	65	help_home_state_climate
18	97	protect_history_continue_security
19	264	security_economic_climate_continue

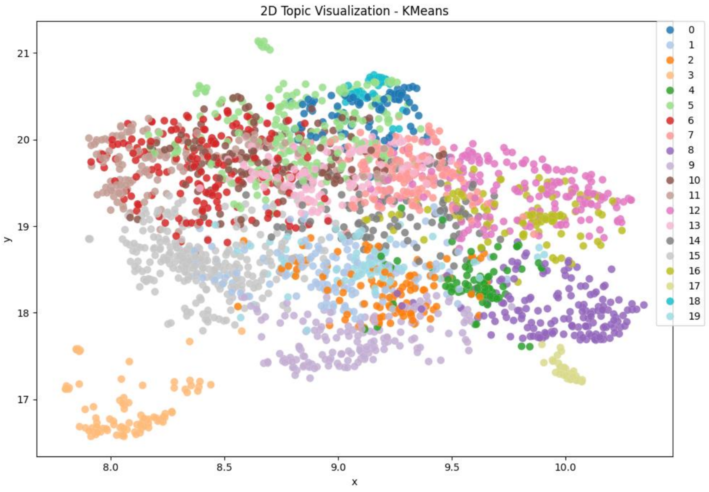

Figure 3. KMeans Plot

3.3.3. NNC Method

The suggested method for Neural Network Clustering resulted in the formation of 20 distinct topic clusters. Interestingly, NNC demonstrated more distinct topic differentiation in specific areas, including clusters that focused on community support, climate investment, and the roles of vice presidents. While there are some overlaps with other models, NNC provided a unique latent representation that improved the differentiation of policy-related discussions. These results are summarized in Table 3, while the topic distribution are visualized in Figure 3.

Table 3. Result of NNC

Topic	Count	Name
0	52	percent_school_month_continue
1	252	pay_folk_percent_big
2	147	vote_vice president_happen_vice
3	109	folk_change_child_home
4	156	life_family_history_member
5	249	security_economic_issue_continue
6	246	freedom_vice president_fight_vice
7	84	security_protect_history_american
8	154	woman_issue_care_vice president
9	154	economy_cost_economic_plan
10	92	school_child_future_community
11	105	security_support_stand_continue
12	119	climate_investment_administration_new
13	140	deal_president_happen_able
14	158	care_child_community_live
15	41	great_history_home_thank thank
16	221	freedom_vice president_vice_history
17	28	vice president_vice_fight_business
18	95	help_state_home_sure
19	145	worker_business_investment_administration

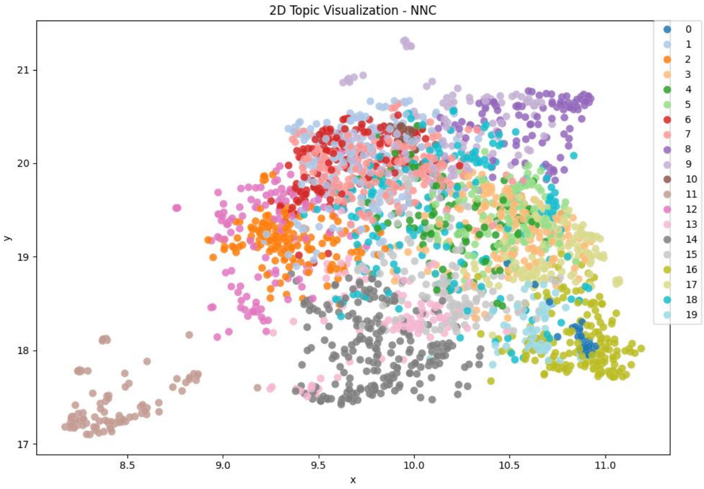

Figure 4. NNC Plot

3.4. Model Evaluation

Model evaluation is an important phase in topic modeling because it allows for an assessment of both clustering quality and topic coherence. This study used four key evaluation metrics: the Davies-Bouldin Index (DBI), UMass, NPMI, and Topic Diversity.

3.4.1. Davies-Bouldin Index (DBI)

The DBI determines the compactness and separation of clusters. A lower DBI score indicates better clustering, as the clusters are both tight and well separated from one another. NNC had the lowest DBI score of the three models, indicating that its clustering structure was the most effective for intra-cluster similarity and inter-cluster dissimilarity. Table 4 summarizes the DBI evaluation results.

Table 4. Evaluation of DBI

Model	Score
HDBSCAN	1.5361
KMeans	1.1737
NNC	1.1674

3.4.2. UMass

The UMass coherence score assesses semantic coherence of topics by calculating the co-occurrence frequency of words within documents. Higher UMass scores (closer to zero) indicate improved topic coherence. All models generated similar UMass scores. NNC performed slightly better than KMeans and HDBSCAN, with scores close to zero indicating moderate coherence. The evaluation results are presented in Table 5.

Table 5. Evaluation of UMass

Model	Score
HDBSCAN	-0.8214
KMeans	-0.8179
NNC	-0.7485

3.4.3. NPMI

NPMI is a semantic-based coherence metric that measures the frequency with which words within a topic co-occur more than by chance. Higher NPMI scores indicate improved semantic consistency. NNC had the highest NPMI score, implying that the topics it generated were more semantically coherent. KMeans followed closely, with HDBSCAN producing slightly lower results. Table 6 shows the results.

Table 6. Evaluation of NPMI

Model	Score
HDBSCAN	0.0916
KMeans	0.0894
NNC	0.0918

3.4.4. Topic Diversity

Topic Diversity assesses the proportion of unique words among all topic keywords, which serves as an indicator of redundancy. Higher diversity scores indicate that the topics include a broader range of vocabulary. NNC once again outperformed the other models, producing topics with the most lexical variety, while HDBSCAN and KMeans had identical scores. The evaluation results are shown in Table 7.

Vol. 6, No. 4, August 2025, Page. 1957-1970
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5090

Table 7. Evaluation of Topic Diversity

Model	Score
HDBSCAN	0.800
KMeans	0.790
NNC	0.815

Figure 5 shows a metric chart comparing the performance of three clustering algorithms, HDBSCAN, KMeans, and NNC, using four assessment metrics: Davies-Bouldin Index (DBI), UMass, NPMI, and Topic Diversity.

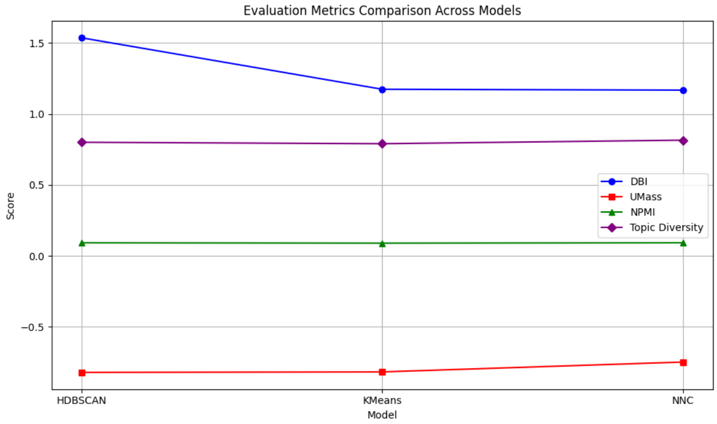

Figure 5. Comparative Metrics Chart

4. DISCUSSIONS

This section presents the results of topic modeling conducted using the BERTopic framework, integrating three distinct clustering algorithms: HDBSCAN, KMeans, and a custom-designed NNC methods. The evaluation focused on each model's ability to generate semantically coherent and diverse topics from a corpus of 2,747 preprocessed presidential speeches delivered by Joe Biden during his tenure.

The HDBSCAN-based model successfully identified several important themes, including freedom, economic policy, and climate administration. However, it generated a large number of outliers, with 787 documents assigned to the noise cluster (Topic -1). This result is consistent with HDBSCAN's nature as a density-based algorithm designed to exclude documents that are not strongly associated with any particular cluster. While this characteristic is useful for detecting ambiguous or weakly related content, it had an impact on overall cluster compactness and separation, as evidenced by the highest Davies-Bouldin Index (DBI) score of 1.5361 among the three models. Semantic coherence was also relatively low, with a UMass score of -0.8214 and an NPMI of 0.0916 indicating moderate consistency. The model's topical diversity score was 0.800. suggesting some redundancy in the topic vocabularies.

In contrast, the KMeans model produced more compact and evenly distributed clusters, with all documents assigned to one of the 20 clusters; no outliers were found. This resulted in greater intra-cluster similarity and inter-cluster separation, as evidenced by a lower DBI score of 1.1737. The model also demonstrated slightly better semantic coherence than HDBSCAN, with a UMass score of -0.8179, though its NPMI value was slightly lower at 0.0894. Topic diversity remained unchanged from HDBSCAN (0.800), indicating that lexical variety did not improve significantly.

Meanwhile, the proposed NNC model produced the best overall results. It had the lowest DBI score (1.1674), indicating the most compact and well-separated clustering structure. In terms of coherence, NNC achieved the best UMass score of -0.7485, indicating greater semantic alignment among topic keywords. Furthermore, NNC achieved the highest NPMI score of 0.0918, indicating stronger word co-occurrence consistency, and outperformed the other models in topic diversity with a score of 0.815, indicating broader lexical coverage.

Finally, the evaluation shows that the clustering algorithm used has a significant impact on the structure, coherence, and lexical richness of BERTopic-generated topics. While each method has its own set of advantages, neural network-based clustering (NNC) consistently outperforms both HDBSCAN and KMeans on key metrics, achieving the lowest DBI, highest coherence (UMass and NPMI), and greatest topic diversity.

This study contributes to BERTopic research by extensively comparing traditional (HDBSCAN, KMeans) and neural clustering methods on a politically significant, large-scale corpus. The results reveals that deep clustering not only improves traditional evaluation metrics but also increases interpretability by providing semantically richer subjects, which is critical for fields such as political discourse analysis, policy monitoring, and media studies. By incorporating NNC into BERTopic, we provide a methodological pathway for improving topic modeling in contexts requiring nuanced semantic differentiation and broad lexical representation, filling a gap in the literature where clustering innovation has received less attention than embedding optimization. This contribution has significance for future political NLP research, as well as any application that requires structural precision and semantic depth in topic extraction.

5. CONCLUSION

This study improves neural clustering-based topic modeling for political text analysis by incorporating a proprietary Neural Network Clustering (NNC) method into the BERTopic framework. The approach was applied to 2,747 official speeches by US President Joe Biden and revealed that clustering decision has a significant impact on topic coherence, lexical diversity, and interpretability. While HDBSCAN and KMeans provided useful baselines for capturing core themes and balanced cluster structures, respectively, NNC consistently generated better findings, with the lowest Davies-Bouldin Index, the highest coherence (UMass and NPMI), and the highest topic diversity.

The primary contribution of this research is to demonstrate that deep learning-driven clustering can significantly improve topic modeling in the political arena, particularly for corpora where small semantic distinctions are critical for interpretation. This study emphasizes the clustering stage as a powerful, underutilized lever for increasing both the structural quality and semantic richness of subjects, pushing beyond embedding optimization. This has obvious implications for computational political science, enabling more reliable tracking of policy narratives, ideological framing, and rhetorical alterations in political discourse.

Future research can extend this work by overcoming its current constraints. One intriguing path is to expand to multilingual corpora, which would allow for the investigation of topic evolution across different languages and cultural situations, opening up new possibilities for global comparative political discourse studies. Another option is to incorporate supervised guidance or domain-specific ontologies into the modeling pipeline, which will drive topic generation toward more task-relevant and interpretable outputs.

Furthermore, improving the neural clustering architecture itself such as by using attention mechanisms or contrastive learning has the potential to sharpen the boundaries between semantically adjacent topics, increasing the granularity and richness of the extracted themes. Following these paths will establish NNC integration within BERTopic as a stable and adaptable process for creating high-fidelity subjects in complicated and high-stakes domains, including political communication and policy analysis.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] R. M. Zain, S. Anggai, Tukiyat, A. Musyafa, and A. A. Waskita, "Revealing a Country ' s Government Discourse Through BERT-based Topic Modeling in the US Presidential Speeches," *2024 Int. Conf. Comput. Control. Informatics its Appl.*, vol. 11, pp. 191–196, 2024, doi: 10.1109/IC3INA64086.2024.10732578.

[2] A. A. Hidayat, R. Nirwantono, A. Budiarto, and B. Pardamean, "BERT-based Topic Modeling Approach for Malaria Research Publication," *2022 Int. Conf. Informatics, Multimedia, Cyber Inf. Syst.*, pp. 326–331, 2022, doi: 10.1109/ICIMCIS56303.2022.10017743.

[3] S. Umamaheswaran, V. Dar, E. Sharma, and J. S. Kurian, "Mapping Climate Themes from 2008-2021 - An Analysis of Business News Using Topic Models," *IEEE Access*, vol. 11, no. March, pp. 26554–26565, 2023, doi: 10.1109/ACCESS.2023.3256530.

[4] D. J. Cahyadi, H. Murfi, Y. Satria, S. Abdullah, and Y. Widyaningsih, "BERT-Based Deep Embedded Clustering for Topic Modeling," *Int. Conf. Comput. Control. Informatics its Appl. IC3INA*, no. 2024, pp. 331–336, 2024, doi: 10.1109/IC3INA64086.2024.10732729.

[5] Y. An, H. Oh, and J. Lee, "Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network," *Appl. Sci.*, vol. 13, no. 16, Aug. 2023, doi: 10.3390/app13169443.

[6] A. S. Kazmali and A. Sayar, "Web Scraping : Legal and Context Ethical Considerations in General Local - A Review," *Procedia Comput. Sci.*, vol. 259, pp. 1563–1572, 2025, doi: 10.1016/j.procs.2025.04.111.

[7] L. R. Nisa, A. Luthfiarta, A. Nugraha, and M. Hasan, "A TOPIC-BASED APPROACH FOR RECOMMENDING UNDERGRADUATE THESIS SUPERVISOR USING LDA WITH COSINE SIMILARITY," *J. Tek. Inform.*, vol. 6, no. 1, pp. 311–323, 2025.

[8] A. Parlina and I. Maryati, "Leveraging BERTopic for the Analysis of Scientific Papers on Seaweed," *2023 Int. Conf. Comput. Control. Informatics its Appl.*, no. 2022, pp. 279–283, 2023, doi: 10.1109/IC3INA60834.2023.10285737.

[9] D. Yohannes, Y. B. Sinshaw, S. H. Asefa, and Y. Assabie, "Amharic document clustering using semantic information from neural word embedding and encyclopedic knowledge," *Sci. African*, vol. 28, p. e02657, 2025, doi: 10.1016/j.sciaf.2025.e02657.

[10] V. Sharifian-attar, J. Li, H. Moss, and J. Johnson, "Analysing Longitudinal Social Science Questionnaires : Topic Modeling with BERT-based Embeddings," *2022 IEEE Int. Conf. Big Data (Big Data)*, pp. 5558–5567, 2022, doi: 10.1109/BigData55660.2022.10020678.

[11] J. Jeoung, J. Hong, J. Choi, and T. Hong, "Analyzing news and research articles about energy storage systems in South Korea based on network analysis and topic modeling," *Energy Build.*, vol. 335, no. January, p. 115547, 2025, doi: 10.1016/j.enbuild.2025.115547.

[12] W. Kang, Y. Kim, H. Kim, and J. Lee, "An Analysis of Research Trends on Language Model using BERTopic," *2023 Congr. Comput. Sci. Comput. Eng. &amp; Appl. Comput.*, no. 2022, pp. 168–172, 2023, doi: 10.1109/CSCE60160.2023.00032.

[13] H. Suryotrisongko, "Topic Modeling for Cyber Threat Intelligence (CTI)," *7th Int. Conf. Informatics Comput.*, pp. 1–7, 2022, doi: 10.1109/ICIC56845.2022.10006988.

[14] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv:2203.05794v1*, 2022.

[15] E. M. Kurniawan, "Analysing Hoax Dataset in Indonesian Language with Topic Modeling," *Int. Conf. ICT Smart Soc.*, vol. 10th, pp. 1–6, 2023, doi: 10.1109/ICISS59129.2023.10291599.

[16] H. Son and Y. E. Park, "Agenda-setting effects for covid-19 vaccination: Insights from 10 million textual data from social media and news articles using BERTopic," *Int. J. Inf. Manage.*, vol. 83, no. April, p. 102907, 2025, doi: 10.1016/j.ijinfomgt.2025.102907.

[17] K. T. Jacob Devlin, Ming-Wei Chang, Kenton Lee, "BERT : Pre-training of Deep Bidirectional

Transformers for Language Understanding," *Proc. ofNAACL-HLT 2019, pages 4171–4186 Minneapolis, Minnesota, June 2 - June 7, 2019. c?2019 Assoc. Comput. Linguist.*, pp. 4171–4186, 2019.

[18] Z. Kastrati, A. L. I. S. Imran, S. M. Daudpota, M. A. Memon, and M. Kastrati, "Soaring Energy Prices : Understanding Public Engagement on Twitter Using Sentiment Analysis and Topic Modeling With Transformers," *IEEE Access*, vol. 11, no. February, pp. 26541–26553, 2023, doi: 10.1109/ACCESS.2023.3257283.

[19] G. Chen, X. Li, Y. Yang, and W. Wang, "Neural Clustering based Visual Representation Learning," pp. 5714–5725, 2024, doi: 10.1109/CVPR52733.2024.00546.

[20] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Expert Syst. Appl.*, vol. 172, no. September 2020, p. 114652, 2021, doi: 10.1016/j.eswa.2021.114652.

[21] G. Tian, P. Wang, R. Wang, and Y. Du, "Smart Contract Classification based on Neural Clustering and Semantic Feature Enhancement," *Blockchain Res. Appl.*, p. 100303, 2025, doi: 10.1016/j.bcra.2025.100303.

[22] M. W. Akram, M. Salman, M. F. Bashir, S. M. S. Salman, T. R. Gadekallu, and A. R. Javed, "A Novel Deep Auto-Encoder Based Linguistics Clustering Model for Social Text," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2022, doi: 10.1145/3527838.

[23] J. P. Lim and H. W. Lauw, "Aligning Human and Computational Coherence Evaluations," *Comput. Linguist.*, no. March, pp. 1–60, 2024, doi: 10.1162/coli_a_00518.

[24] A. Bachoumis, C. Mylonas, K. Plakas, M. Birbas, and A. Birbas, "Data-Driven Analytics for Reliability in the Buildings-to-Grid Integrated System Framework : A Systematic Text-Mining-Assisted Literature Review and Trend Analysis," *IEEE Access*, vol. 11, no. October, pp. 130763–130787, 2023, doi: 10.1109/ACCESS.2023.3335191.

[25] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," *Proc. Ger. Soc. Comput. Linguist. (GSCL 2009)*, pp. 31–40, 2009.

[26] M. Białas, M. M. Mirończuk, and J. Mańdziuk, "Leveraging spiking neural networks for topic modeling," *Neural Networks*, vol. 178, no. May, p. 106494, 2024, doi: 10.1016/j.neunet.2024.106494.

[27] J. Song, Y. Yuan, K. Chang, B. Xu, J. Xuan, and W. Pang, "Exploring public attention in the circular economy through Topic Modeling with twin hyperparameter optimisation," *Energy AI*, vol. 18, no. May, p. 100433, 2024, doi: 10.1016/j.egyai.2024.100433.