Graph Regularized Probabilistic Latent Semantic Analysis for Topic Analysis Using Social Media Data

Muhammad Panji Muslim[1)*], Novi Trisman Hadi[2)], Muhammad Adrezo[3)]
[1)2)3)]Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
[1)]muhammadpanji@upnvj.ac.id, [2)]novitrismanhadi@upnvj.ac.id, [3)]muhammad.adrezo@upnvj.ac.id

Submitted :Dec 16, 2024 | **Accepted** : Jan 14, 2024 | **Published** : Jan 17, 2025

Abstract: In today's digital era, social media data provides valuable insights into public opinion. This study implements the Graph Regularized Probabilistic Latent Semantic Analysis (GPLSA) method to analyze topics from social media data surrounding the 2024 Indonesian Presidential Election (Pemilu), as well as to evaluate the efficiency of the Probabilistic Latent Semantic Analysis (PLSA) algorithm. The research stages include collecting social media data on presidential debates and elections, text pre-processing, and applying the GPLSA method to identify main topics. The analysis results show that PLSA without graph achieved a topic coherence score of 0.653, indicating good consistency, while GPLSA decreased to 0.5, suggesting that the addition of graph regularization did not significantly enhance coherence. Additionally, PLSA without graph achieved a perplexity score of 12.138, indicating good predictive capability, while GPLSA increased to 12.511, showing that graph regularization did not improve the prediction of new words. PLSA without graph also produced topics relevant to election issues, while GPLSA generated topics influenced by graph regularization, though without significant improvement in topic quality. Sentiment analysis of social media posts provides insights into public responses to debates and election issues. Validation of the GPLSA model ensures relevant topic representation. This research contributes to the development of text analysis methods and offers valuable information for elections and democratic participation. These results can be utilized by stakeholders to make more strategic and informed decisions.

Keywords: Probabilistic Latent Semantic Analysis; Graph Regularized; Topic Analysis; Media Social; Indonesian Presidential Election;

INTRODUCTION

Currently, social media is the most popular platform for long-distance communication, with millions of users globally sharing thoughts, experiences, and opinions. The rapid growth of social media has made people more active in sharing everything happening around them, including the current focus on democratic politics leading up to the 2024 Presidential Election in Indonesia. This trend has led to a dynamic exchange of public reviews and comments about the presidential candidates, as people express their preferences and opinions on various platforms (Arianto, 2021; Juliswara & Muryanto, 2022). The political discourse on social media can significantly influence voters' perceptions and actions, which has been demonstrated in previous studies of political elections, such as the U.S. 2020 presidential election, where social media was pivotal in shaping voter behavior and engagement (Tufekci, 2020).

In the digital transformation era, social media has become a primary source for individuals to share opinions, views, and up-to-date information, especially in the political context. Analyzing data from social media has become essential for understanding the dynamics of public opinion, particularly in anticipation of significant events such as the presidential candidate debates (Onah, Pang, & El-Haj, 2022; Krishnan, 2023). Research has shown that social media analysis, especially of political conversations, can uncover sentiment shifts and public opinion trends (Bodo, 2022). These insights are crucial for understanding voter behavior and predicting election outcomes, as evidenced by studies analyzing Twitter data during the 2016 U.S. election (Tumasjan et al., 2010).

Topic modeling is a technique that generates document representations through keywords, which are then used in the indexing and retrieval processes to match user needs. Latent Semantic Analysis (LSA) was the first method to produce document representations in terms of word groups, using the Bag-of-Words approach (Zhang,

*name of corresponding author

Liu, & Yan, 2023). Building on LSA, Probabilistic Latent Semantic Analysis (PLSA) assigns topic weights to each document based on probabilistic values, allowing for deeper insights into hidden topic structures (Kharisudin & Masri'an, 2021). PLSA has been applied in various domains, including sentiment analysis and news classification, to identify underlying topics from large datasets (Monika et al., 2024). Furthermore, Graph Latent Semantic Analysis (GLSA) modifies the term-document matrix into n-grams, introducing a more sophisticated way of mapping term relationships (S, Findawati, & Indahyanti, 2023). Recent advancements in topic modeling include the development of multidimensional techniques such as Multidimensional Latent Semantic Analysis (MDLSA) and Syntactically Enhanced Latent Semantic Analysis (SELSA), which offer additional refinements by exploring term relationships, spatial distributions, and syntactic aspects (Shiju, 2022; Griciūtė, Han, & Nenadic, 2023).

Topic modeling remains a vast field with applications across multiple domains. In this research, we specifically focus on PLSA as one of the most prominent topic modeling techniques. PLSA excels in uncovering hidden topic structures within text data by leveraging the probability distribution of words. By analyzing these structures, PLSA provides a method for probabilistic topic analysis in text, uncovering deeper semantic patterns in large-scale datasets (Grootendorst, 2022). Recent advances have introduced Graph Regularized PLSA (GPLSA), where data entities are mapped into an undirected graph, and the similarity between topic compositions is measured by the divergence between discrete probabilities (Dieng, Ruiz, & Blei, 2020). This method extends the original PLSA algorithm by integrating graph regularization and can be adapted for multi-modal data by establishing connections between data entities from different sources (Isoaho, Gritsenko, & Mäkelä, 2021).

LITERATURE REVIEW

A literature study is conducted for this research to ensure that the research methodology design is clear and structured. The literature review will focus on studies related to the three clustering methodologies to be used: Social Media, the Hierarchical model, and the Gaussian Mixture model.

Social Media, Presidential Election

The presidential election in Indonesia has undergone significant evolution over the past decade, especially with the emergence and dominance of social media as a campaign tool. Social media, with millions of active users in Indonesia, has become the primary battleground for presidential candidates to influence voters. The COVID-19 pandemic accelerated the transformation of digital culture in Indonesia.

In addition, digital literacy has become crucial in helping voters understand and filter the information they receive. In presidential elections, incorrect or misleading information and topics can significantly impact voter perceptions and election outcomes. One of the main challenges in presidential elections is the spread of false information or hoaxes. In the context of presidential elections, a similar approach can be applied to ensure that voters receive accurate and reliable information.

In conclusion, presidential elections in Indonesia have been significantly influenced by the emergence of social media. This study utilizes user comments related to the 2024 presidential election candidates up to the time of the democratic event. These comments serve as topic analysis material for the research model used in this study.

Topic Modeling

Topic Modeling is an unsupervised machine learning method that applies clustering to discover latent variables within large text data. Topic modeling is a technique used to find document representations in the form of keywords extracted from documents. These keywords are then used for indexing and retrieving documents as per the user's needs.

Topic modeling comprises a series of algorithms aimed at identifying and describing documents with thematic information, associating various themes with entities based on unified learning through themes. There are four main methods of text data topic modeling, as cited from.

a. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method or technique in the field of Natural Language Processing. The primary goal of LSA is to create a vector-based representation to compare the semantic similarity of different words or documents.

b. Probabilistic Latent Semantic Analysis (PLSA)

Probabilistic Latent Semantic Analysis (PLSA) is an approach introduced to address certain limitations found in LSA. PLSA automates document indexing based on a statistical latent class model for factor analysis of large data sets and aims to enhance Latent Semantic Analysis by using a probabilistic generative model.

c. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a widely used text mining algorithm based on a statistical (Bayesian) topic model. LDA is a generative model that attempts to simulate the writing process and generate documents on given topics.

d. Correlated Topic Model (CTM)

Correlated Topic Model (CTM) is a type of statistical model used in Natural Language Processing and Machine Learning. CTM is used to find topics presented within a group of documents. The key to CTM is the logistic normal distribution. CTM depends on LDA for its foundation.

Probabilistic Latent Semantic Analysis

Before the introduction of Probabilistic Latent Semantic Analysis (PLSA) in 1999, Latent Semantic Analysis (LSA) was developed by Landauer, Foltz, and Laham in 1998. LSA is an automated statistical technique used to compare the semantic similarity of words or documents by uncovering the meaning or concepts behind them, rather than focusing on syntax or style. It maps words or documents into a "concept space" (latent semantic space) by reducing a high-dimensional matrix into a lower dimension, while still representing the document's content. Singular Value Decomposition (SVD) is used in LSA to decompose the matrix and measure similarity. PLSA, which builds on the aspect model, aims to generate effective document representations by linking documents to words through key keywords, known as aspect models. These models assume conditional independence between documents and words, connected by topics. Documents and words under conditional independence are connected by topics, as illustrated in Figure 1 below:

Documents
(D)
Topics
(Z)
Words
(W)

Fig. 1 Relationship between Document, Topic, and Word

The joint probability between a document (D) and a word (W) is represented in the following equation:

$$P(d,\omega) =\sum_{z\epsilon Z} P(z)\, P(d \mid z)\, P(\omega|z) \tag{1}$$

PLSA uses probability matrices to model the relationship between documents (D), words (W), and topics (Z), differing from LSA's term-frequency approach. PLSA matrices represent the likelihood of topics associated with documents and words, introducing conditional independence where topics (Z) mediate the relationship between W and D. The model starts with a stochastic process or greedy initialization to generate initial probability values, which are then refined through iterative training using the Expectation Maximization (EM) algorithm.

The EM algorithm aims to minimize error values by optimizing weights, which reduces error with each iteration. As the number of training iterations increases, topics and probabilities become more defined. The matrices resulting from PLSA include P(d|z), which represents topic distribution across documents, P(w|z), showing topic distribution across words, and P(z), reflecting the probabilities of the topics themselves.

PLSA improves computational efficiency by addressing the high dimensionality of term-document matrices through dimensionality reduction. Instead of directly working with a large matrix (e.g., 100x1000, containing 100,000 entries for 100 documents and 1,000 words), PLSA introduces latent topics (Z) and decomposes the distribution into smaller matrices. For 10 topics, this reduces the matrix dimensions to 100x10 and 1000x10, resulting in only 11,000 entries. Using methods like Singular Value Decomposition (SVD), PLSA maintains accuracy while significantly enhancing efficiency.

Graph Regularized Probabilistic Latent Semantic Analysis (GPLSA)

This study begins with the general setup of the single-model GPLSA problem and extends the formulation to the multi-modality problem. At this stage, refer to the examples in Figures 3 (a) and (b).

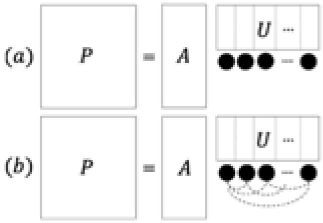


Fig. 2 (a) PLSA: vertices (columns of U) have no constraints, (b) GPLSA: vertices with a graph regularizer.

*name of corresponding author

This study presents an extension of the Probabilistic Latent Semantic Analysis (PLSA) to a multi-modality problem, introducing Graph Regularized PLSA (GPLSA). The study formulates GPLSA as a constrained optimization problem with a graph regularizer that helps smooth the effects of similar semantic entities. Specifically, the researchers implement the GPLSA algorithm as a constrained optimization problem for the graph regularizer

It defines the optimization function to minimize, which consists of a standard PLSA term and a co-regularization term. The optimization process alternates between updating matrices A and U using block coordinate descent, with the EM algorithm applied in each step.

The optimization of A follows the standard PLSA method, while the optimization of U incorporates a graph regularizer. The study provides an auxiliary function to derive a lower bound for U and iterates the optimization using a specific update rule. The final goal is to optimize the objective function, and the study applies different types of divergences (KL, l2, l1) as regularizers, explaining the solution methods for each type. The l1 divergence step is simpler and more efficient than the others, and it can lead to U components being identical, which does not happen with l2 or symmetric KL regularizers.

METHOD

This study develops an effective topic analysis algorithm by implementing Probabilistic Latent Semantic Analysis (PLSA) enhanced with a graph-based regularization mechanism (Graph-Regularized PLSA/GPLSA). In GPLSA, data entities are mapped onto an undirected graph, and the similarity between topic compositions is measured by the divergence between discrete probabilities, which serves as the graph regularizer. The GPLSA algorithm is extended to handle multiple data modalities, with three popular regularizers (l1, l2, and symmetric KL divergence) used in the iterative algorithm. The dataset for this study is collected from social media using crawling or API-based methods, with text documents preprocessed through cleaning, tokenization, and removal of stop words. The PLSA algorithm is initialized with random probability distributions, and graph-based regularization is applied to incorporate structural information into the model, ensuring related topics share similar probability distributions. The developed algorithm consists of the following phases:

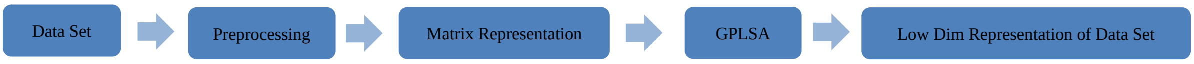


Fig. 3 *Graph Regularized Probabilistic Latent Semantic Analysis Algortihm*

RESULT

This section will outline the steps involved in implementing the Probabilistic Latent Semantic Analysis algorithm with Graph Regularization for topic analysis using social media data related to the 2024 Indonesian Presidential Election. This chapter also describes the testing scenarios and the evaluation of the proposed method through the analysis of the test results. The findings from this analysis will be presented at the end of the section.

Data Preparation Stages

The data preparation stage is essential to ensure that collected data is relevant, clean, and ready for analysis. In this study, Twitter data was gathered using the Basic Subscription API package, with the preparation process involving multiple steps to ensure the data's suitability for analysis:

1) **API Registration and Configuration**

   In this stage, the registration and configuration of the API are conducted to access data from Twitter. This process involves registering on the Twitter Developer Platform, obtaining API credentials, and configuring the API in the code.

2) **API Configuration in the Source Code**

   The next step is to configure the API credentials in the code to collect data according to the research requirements.

3) **Data Collection**

   At this stage, the researcher collects data from Twitter based on the specified criteria and then stores it in a format suitable for further analysis, as configured during the data preparation stage. Below is a detailed explanation and the technical steps involved:

   a. **Defining Search Criteria**

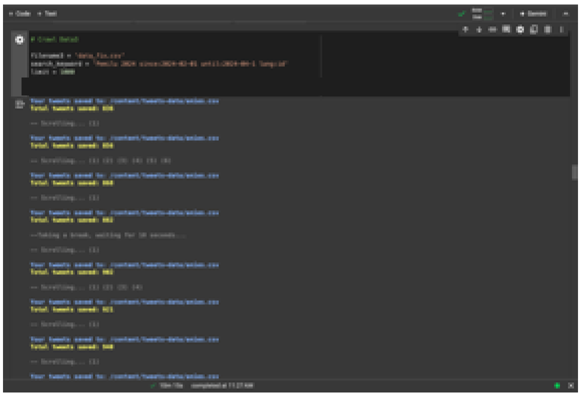
      The first step in data collection is to define search criteria, including selecting relevant keywords, hashtags, or phrases, and setting parameters to filter the results. For this study, the keywords “Pemilu 2024, Prabowo, Gibran, Anies, Ganjar, Mahfud” were chosen, with a crawling time frame from February 1 to April 1, 2024. It is important to select specific keywords to ensure the

*name of corresponding author

collected data is relevant to the research topic, as overly general keywords may produce irrelevant data, while overly specific ones may yield insufficient data.

b. Fetching Data with API

At this stage, the Twitter API is utilized to collect tweets based on the predefined search criteria. This process involves using API methods to perform searches and gather relevant data. The data collection process using the API and the configured keywords is illustrated in Figure 4.



```
# Crawl Data

filename = 'data_fix.csv'
search_keyword = 'Pemilu 2024 since:2024-02-01 until:2024-04-1 lang:id'
limit = 1000
```

```
Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1) (2) (3) (4) (5) (6)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

--Taking a break, waiting for 10 seconds...

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1) (2) (3) (4)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/anies.csv
Total tweets saved: [illegible]

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/anies.csv
```

Fig. 4 Fetching Data with the API.

The Twitter API provides methods for searching tweets based on specified criteria. One commonly used method is *tweepy.Cursor*, which simplifies the retrieval of large amounts of data. The Cursor allows for incremental searches and the collection of tweets that match your search criteria. This is particularly useful for handling pagination and ensuring comprehensive results.

c. Storing Data

Once the tweet collection process is complete, the raw data is stored in a format suitable for analysis, ensuring easy access and management for future use. The collected data is processed and organized into a structure like a CSV file, which is commonly used due to its simplicity and compatibility with various data analysis tools such as Excel or pandas in Python. After organizing the data into the desired format, it is stored either locally or in the cloud to support efficient handling of large datasets and accelerate computational processes. This study utilizes Google Cloud with an additional 2 TB subscription, ensuring smooth analysis without local resource limitations and enabling faster, more accurate research outcomes. Secure and private data storage is crucial, especially for sensitive information, requiring restricted access and periodic backups to prevent data loss. Properly stored data facilitates subsequent analysis, such as visualization or algorithm application. In this research, approximately 17,565 tweets related to the 2024 Indonesian Presidential Election were collected using the API Subscribe Basic configuration. This data includes various opinions, discussions, and information related to the topic, which will then be used for further analysis in the study

Preprocessing Data

Data preprocessing is a crucial step to ensure high-quality analysis of 17,565 tweets collected from the X platform. It involves cleaning the data by removing irrelevant elements like URLs, mentions, hashtags, and special characters using regular expressions. The text is then normalized to lowercase, tokenized into words, and stopwords are eliminated to focus on meaningful terms. Stemming tools like Sastrawi are used to reduce words to their root forms, simplifying analysis. The cleaned data is converted into numerical formats, such as bag-of-words, to be processed by the modified Probabilistic Latent Semantic Analysis (PLSA) algorithm with Graph Regularization. This thorough preprocessing ensures accurate and meaningful analysis results.

Algorithm Development Stages

The Algorithm Development phase of this study outlines the methods and technical processes used to create and enhance the Graph Regularization-modified Probabilistic Latent Semantic Analysis (PLSA) algorithm for topic analysis on social media data related to the 2024 Indonesian Presidential Election. The process involves three key stages:

1) Basic PLSA Implementation

This stage implements the foundational PLSA model using preprocessed tweet data, represented as a document-word matrix. Key steps include Model Initialization, Randomly initialize topic distributions for documents P(z|d) and word distributions for topics P(w|z). EM Algorithm, Use the Expectation-Maximization algorithm:

- E-step: Calculate posterior probabilities of topics for document-word pairs.
- M-step: Update model parameters to maximize log-likelihood.

Convergence, Iterate until convergence to produce topic distributions for documents and word distributions for topics. The result identifies main topics in tweets about the election and their distribution, forming the foundation for applying Graph Regularization. The model generates 20 topics, reflecting the complexity of the data.

2) Graph Regularization Integration

In this stage, Graph Regularization enhances the PLSA model by incorporating structural information from a graph into the algorithm's objective function. Key steps include Adding a graph-based regularization term to the PLSA objective function, Adjusting topic distributions based on the graph's structural relationships between words or documents. This integration refines the topic distributions to account for both graphical connections and the PLSA model's topic information, resulting in a more accurate representation of topics.

3) Algorithm Testing and Validation

The final stage evaluates the model's performance using:

- **Coherence Scores:** Measures how interpretable and meaningful the topics are by assessing the co-occurrence of words within the same topic.
- **Perplexity:** Indicates the model's ability to predict unseen data, serving as a measure of its generalization capability.

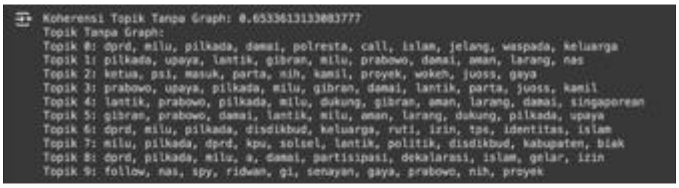
These metrics validate the model's ability to generate coherent and predictive topics for analyzing election-related tweets.

Implementation and Analysis Stage

a. Comparison of the Implementation of PLSA Algorithm with and without Graph Regularization

1. PLSA Model without Graph Regularization

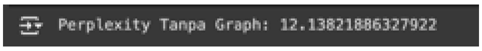
In this stage, the PLSA model without graph regularization is initialized using processed corpus data and trained with specified parameters, such as the number of topics and iterations. This process aims to identify a topic representation based on probabilistic techniques. The resulting topic distributions show how words are grouped into topics, reflecting frequently occurring word pairs in the text, but without graph-based information. The topic coherence value, calculated from the model's testing, measures the consistency of word co-occurrence within topics. A coherence value of 0.653 indicates a relatively good level of cohesion, suggesting meaningful topics without graph regularization. As shown in Figure 5 below:



```
Koherensi Topik Tanpa Graph: 0.6533613133083777
Topik Tanpa Graph:
Topik 0: dprd, milu, pilkada, damai, polresta, call, islam, jelang, waspada, keluarga
Topik 1: pilkada, upaya, lantik, gibran, milu, prabowo, damai, aman, larang, nas
Topik 2: ketua, psi, masuk, parta, nih, kamil, proyek, wokeh, juoss, gaya
Topik 3: prabowo, upaya, pilkada, milu, gibran, damai, lantik, parta, juoss, kamil
Topik 4: lantik, prabowo, pilkada, milu, dukung, gibran, aman, larang, damai, singaporean
Topik 5: gibran, prabowo, damai, lantik, milu, aman, larang, dukung, pilkada, upaya
Topik 6: dprd, milu, pilkada, disdikbud, keluarga, ruti, izin, tps, identitas, islam
Topik 7: milu, pilkada, dprd, kpu, solsel, lantik, politik, disdikbud, kabupaten, biak
Topik 8: dprd, pilkada, milu, a, damai, partisipasi, dekalarasi, islam, gelar, izin
Topik 9: follow, nas, spy, ridwan, gi, senayan, gaya, prabowo, nih, proyek
```

Fig. 5 Network Graph of Words.

The generated topics tend to include words directly related to the theme of elections and politics, such as "prabowo," "pilkada," "damai," and "gibran." This indicates that the model is capable of identifying and grouping semantically relevant words in the desired context. Finally, we calculate the perplexity of the PLSA model without the graph, with results shown in Figure 6 below:

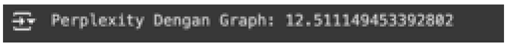


```
Perplexity Tanpa Graph: 12.13821886327922
```

Fig. 6 Perplexity Without Graph

2. PLSA Algorithm with Graph Regularization

The PLSA model with graph regularization was initialized similarly to the standard PLSA, but with an added graph-based regularization to improve topic coherence and inter-topic connections by incorporating structural information from the graph. This adjustment required parameter modifications in the PLSA With Graph model. The results showed that topic distributions reflected word groupings with enhanced semantic relationships, as contextually relevant terms became more prominent. However, after calculating topic coherence, the topics displayed a broader word distribution and lower consistency related to the election theme, suggesting increased diversity or altered word clustering. The perplexity score of 12.511 indicated a slight decline in the model's predictive ability, implying that the graph regularization may not have improved performance or may have introduced noise. This is further illustrated in Figure 7:



```
Perplexity Dengan Graph: 12.511149453392802
```

Fig. 7 Perplexity Score of PLSA with Graph Regularization.

DISCUSSIONS

The implementation of the Graph Regularized Probabilistic Latent Semantic Analysis (GPLSA) algorithm for analyzing topics related to the 2024 Indonesian Presidential Election revealed several key findings. PLSA without graph regularization achieved a higher topic coherence score (0.653) and lower perplexity (12.138) compared to PLSA with graph regularization, which scored 0.5 and 12.511, respectively.

Analysis of Results

The results of the experiments are summarized in Table 1. The data presented reflects the outcomes of testing the PLSA implementation both without Graph Regularization and with the inclusion of Graph Regularization.

a. Topic Coherence

In these results, PLSA without graph regularization shows a higher coherence value (0.653) compared to PLSA with graph regularization (0.5). This indicates that the topics generated without graph regularization are more cohesive and have better word consistency. These topics tend to be clearer in representing specific themes, as the words that frequently appear together are in the same context.

On the other hand, the lower coherence value for the model with graph suggests that the integration of graph information did not significantly improve the cohesion of the topics. This may be because graph regularization did not effectively contribute to enhancing the semantic relationships between the words within the topics. A comparison of the topic coherence values is illustrated in Figure 8 below.

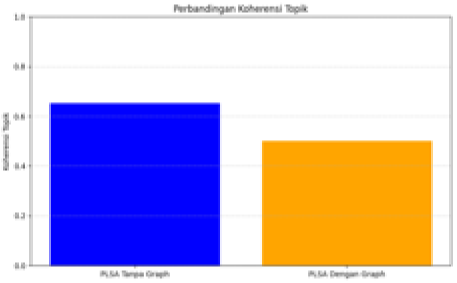


Fig. 8 Topic Coherence Comparison.

b. Topic Perplexity

Perplexity measures how well a model predicts unseen words, with lower values indicating better predictive performance and higher accuracy in handling new data.

- **PLSA Without Graph Regularization**, the perplexity score of 12.138 signifies better performance in predicting unseen words. This result implies that the model without graph regularization effectively captures the relationships in the data and generalizes well to new instances.
- **PLSA With Graph Regularization**, the perplexity score of 12.511 is slightly higher, indicating a decline in the model's ability to predict unseen words accurately. This suggests that the graph regularization component might not have added substantial value or may have introduced additional complexity that the model struggled to handle.

This comparison reveals that PLSA without graph regularization performs more effectively in terms of predictive accuracy. The regularization did not enhance the model's capability for unseen data, possibly due to insufficient or noisy graph information. The comparison of topic perplexity values is illustrated in Figure 9.

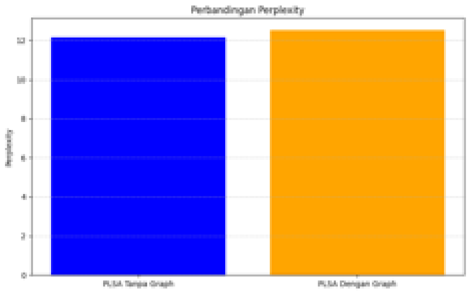


Fig. 9 Topic Perplexity Comparison.

c. Topic Distribution

The topic distribution generated by the model without a graph shows clear clustering of words based on their frequency of occurrence in the data. Words within these topics tend to exhibit a direct and consistent relationship. The Word Cloud of Topics for the topic distribution results from PLSA without a graph is shown in Figure 10.

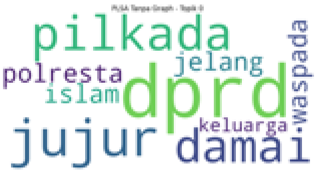


Fig. 10 Word Cloud of Topics from PLSA topic distribution without a graph

Meanwhile, the topic distribution from the model with a graph demonstrates the additional influence of graph regularization. Although this approach attempts to enhance the relationships between words, the results may not provide a better understanding of the topics compared to the model without a graph. Words in topics with a graph might be more influenced by the structural relationships introduced through the graph.

Below is a summary of the analysis and comparison results between the implementation of PLSA without Graph Regularization and with Graph Regularization, as summarized in Table 3.

Table 3. Comparative Testing Results of PLSA without and with Graph Regularization

Evaluation	PLSA Without Graph	PLSA With Graph
Topic Coherence	0.653	0.5
Perplexity	12.138	12.511
Topics	Topics with frequently co-occurring words without graph regularization	Topics influenced more by graph information
Topic Distribution	Word probabilities per topic without graph regularization	Word probabilities per topic with graph regularization

These results indicate that graph regularization did not enhance topic coherence or predictive performance, as the additional constraints imposed by the graph structure may not align well with the natural co-occurrence patterns in social media data. Without graph regularization, the PLSA model more effectively captured the inherent structure of the data, generating topics that were clearer and more relevant to political discourse. In contrast, topics produced with graph regularization appeared less coherent, possibly due to noise or inconsistencies introduced by the graph's influence. This study underscores a limitation in applying graph-based techniques to social media data, highlighting that natural word co-occurrence may hold greater value in such contexts. While the findings align with studies that emphasize the challenges of integrating graph structures in topic modeling, the results suggest the need for more tailored or sophisticated approaches to graph regularization. A notable limitation is the inability of the current graph structure to fully align with the semantic patterns of the dataset, presenting a threat to the validity of applying this technique in similar analyses. Future research could address this by exploring alternative graph-based methods or hybrid models to improve the accuracy and coherence of topic analysis in political discourse.

CONCLUSION

This study concludes that the PLSA model without Graph Regularization consistently outperforms the version with Graph Regularization. This is evidenced by a higher topic coherence score (0.653 vs. 0.5) and a lower perplexity value (12.138 vs. 12.511), indicating the superior ability of the model without graph regularization to generate topics that are semantically consistent and relevant to the data. The integration of Graph Regularization did not provide significant improvements in topic quality or predictive accuracy, demonstrating that the PLSA model without graph regularization is more effective for topic analysis in this dataset.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Research and Community Service Institution (LPPM) of Universitas Pembangunan Nasional Veteran Jakarta for the financial and administrative support provided for this research. This work was conducted under the RISTA (Research based on SINTA). We also thank the members of LPPM and other colleagues who provided insights and expertise that greatly assisted the research. Their valuable feedback and suggestions during the preparation of this study are highly appreciated.

REFERENCES

Arianto, B. (2021). *Pandemi Covid-19 dan Transformasi Budaya Digital di Indonesia*. Titian Ilmu: a. Jurnal Ilmiah Multi Sciences, 5 (2), 45-56.

Bodo, B. (2022). *Social media and political engagement: Understanding the dynamics of political discourse on social platforms*. Journal of Digital Politics, 8(4), 231-247.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453.

Griciūtė, B., Han, L., & Nenadic, G. (2023). Topic modelling of Swedish newspaper articles about coronavirus: A case study using latent Dirichlet allocation method. *arXiv preprint arXiv:2301.03029.*

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based tf-idf procedure.

International Journal of Advanced Computer Science and Applications (IJACSA). (2020). Crime Data Analysis Methodologies for Digital Forensics on Twitter. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 1-11.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal,* 49(1), 300–324.

Juliswara, V. and Muryanto, F. (2022). Model Penanggulangan Hoax Mengenai Berita Covid 19 untuk Pengembangan Literasi Digital Masyarakat di Indonesia. *Jurnal Ilmiah Ilmu Pendidikan*, 5 (7), 80-90.

Kharisudin, I., & Masri'an, H. (2021). Topic modeling on WhatsApp user reviews using latent Dirichlet allocation. *Scientific Journal of Informatics,* 9(1), 1–10.

Krishnan, A. (2023). Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis. *arXiv preprint arXiv:2308.11520.*

Monika, W., Amelia, V., Aris, Q. I., & Nasution, A. H. (2024). Topic modeling of Indonesian children's literature using latent semantic analysis. In *Proceedings of the 2nd International Conference on Environmental, Energy, and Earth Science (ICEEES 2023),* 1–6.

Onah, D. F. O., Pang, E. L. L., & El-Haj, M. (2022). A data-driven latent semantic analysis for automatic text summarization using LDA topic modelling. *arXiv preprint arXiv:2207.14687.*

Shiju, A. (2022). Covid-19 Tweets Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling. *Florida State University's Undergraduate Research Journal*, 12(1), 8-27.

S, U. M., Findawati, Y., & Indahyanti, U. (2023). Topic modeling in COVID-19 vaccination refusal cases using latent Dirichlet allocation and latent semantic analysis. *Jurnal Teknik Informatika (Jutif),* 4(5), 951–960.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International Conference on Weblogs and Social Media, 178-185.

Tufekci, Z. (2020). Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press.

Zhang, L., Liu, J., & Yan, Q. (2023). Graph2topic: An open-source topic modeling framework based on sentence embedding and community detection. *arXiv preprint arXiv:2304.06653.*

Zihan, Z., Fang, M., Chen, L., & Namazi Rad, M. R. (2022). Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3886–3893.