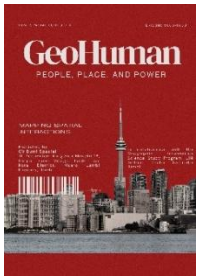
GeoHuman: People, Place, and Power
Vol. 01 No. 01
e-ISSN: ----- ----
DOI: https://doi.org/10.30631/geohuman.v5i1.3223
https://geohuman.bumi-spasial.com/index.php/Journal/index



Mapping Public Opinion Topics on Educational Policy on Twitter Based on Geospatial Aspects Using Latent Dirichlet Allocation (LDA)

Muharoma Fahnur Ihsandi[1], Shalsa Rhamadani[2], Urwawuska Ladini[1]

[1]Statistics Department, Faculty of Science and Technology, Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi
[2]Tadris Mathematics (Mathematics Education), Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi

ARTICLE INFO

Research Article

Keywords: *text mining, topic modelling, latent Dirichlet allocation, public opinion, education policy*

Received: 20-12-2025
Accepted: 25-02-2026
Published: 28-02-2026

ABSTRACT

Educational policy has emerged as a prominent public issue that is widely discussed across social media platforms. As an open-access microblogging platform, Twitter generates a substantial volume of user-generated content that reflects public opinion in real time. Such data provide valuable insights for understanding societal responses to government policies. This study aims to examine the topic segmentation of public opinion regarding educational policies in Indonesia by employing the Latent Dirichlet Allocation (LDA) model. The dataset comprised 8,030 Indonesian-language tweets collected using education-related keywords. After a relevance filtering process, 699 tweets were retained for analysis. The text preprocessing procedures included case folding, removal of numerical characters and punctuation marks, elimination of Indonesian stopwords, and stemming to normalize word forms. The cleaned corpus was then transformed into a Document–Term Matrix (DTM) representation prior to topic modeling. LDA was applied with three predefined topics to extract latent thematic structures within the dataset. The results reveal that public discourse can be categorized into three principal themes: (1) government policies and national conditions, (2) government performance and public policy implementation, and (3) higher education issues related to university students. The topic distribution indicates that discussions concerning government policy and higher education issues are the most dominant themes within public conversations on Twitter. These findings contribute to the growing body of research on social media analytics in public policy studies and provide empirical evidence that may assist policymakers in identifying public concerns and evaluating policy communication strategies.

**Corresponding author:* muharomafahnurihsandi@gmail.com || *Bumi Spasial Publisher*

Introduction

The rapid development of social media has significantly transformed communication patterns in how individuals express opinions, aspirations, and criticisms regarding public policies, including educational policies. Twitter has emerged as an active platform for public discourse due to its real-time nature, openness, and ability to reach diverse segments of society. Within the educational context, Twitter is frequently utilized by students, academics, educators, and the general public to discuss issues such as access to education, tuition fees, learning quality, and government policies in the education sector (Mujahid et al., 2021; Sun & Yan, 2023).

The vast volume of textual data generated on social media presents substantial challenges for manual public opinion analysis. Consequently, text mining and Natural Language Processing (NLP) techniques have become relevant approaches for extracting latent information from unstructured textual data. One of the most widely used topic modeling methods is Latent Dirichlet Allocation (LDA), which enables the identification of hidden thematic structures within a corpus by analyzing word co-occurrence patterns without requiring prior labeling (Negara & Triadi, 2022).

Previous studies have demonstrated the effectiveness of LDA in identifying dominant themes in Twitter-based public discourse. Mujahid et al. (2021) applied LDA to examine online education discussions during the COVID-19 pandemic and identified several key issues reflecting societal responses to changes in the educational system. Other studies have confirmed that LDA-based topic modeling can effectively capture public attitudes toward higher education, academic regulations, and governmental roles in the education sector (Göçen et al., 2024; Khotimah et al., 2025).

However, in the Indonesian context, research focusing specifically on the systematic segmentation of educational policy topics using social media data remains limited. While several studies have applied LDA to explore national issues discussed on Twitter, in-depth analysis of educational policy discourse is still insufficient (Patmawati & Yusuf, 2021; Arianto & Anuraga, 2025). A clearer understanding of thematic structures in public opinion is essential for evaluating policy effectiveness and developing more responsive educational strategies. Therefore, this study aims to analyze the topic segmentation of public opinion regarding educational policies in Indonesia using the Latent Dirichlet Allocation (LDA) method based on Twitter data. By identifying dominant themes emerging in public discourse, this research seeks to provide a comprehensive overview of educational issues that receive significant public attention and to offer empirical insights for policymakers and stakeholders in the education sector.

Methodology

This study employed an exploratory quantitative research design utilizing text mining and topic modeling techniques to examine the segmentation of public opinion regarding educational policies in Indonesia. The dataset was obtained from Twitter and comprised 8,030 Indonesian-language tweets. The data were structured in a data frame containing the original text, cleaned text, sentiment labels, and available location metadata. The primary analysis was conducted on the *final_text* variable, as it had undergone preliminary normalization and was deemed suitable for computational processing. An initial filtering procedure was performed to identify tweets relevant to educational issues using a keyword-based approach. The selected keywords encompassed both formal and non-formal educational terminology, including *education, school, teacher, curriculum, tuition fees (UKT), scholarship, college, student, lecturer,* and *campus*. Filtering was conducted using case-insensitive regular expression pattern matching to ensure comprehensive retrieval. This process resulted in 699

tweets explicitly related to educational discourse, which constituted the analytical corpus of the study.

Subsequently, the textual data underwent systematic preprocessing to enhance analytical validity and reliability. The preprocessing stages included text normalization through lowercase transformation, removal of numerical characters and punctuation, elimination of Indonesian stopwords using a manually curated stopword list, removal of excessive whitespace, and stemming to reduce lexical variations to their root forms. All preprocessing procedures were implemented using the tm and SnowballC packages within the R statistical computing environment. The cleaned corpus was transformed into a Document–Term Matrix (DTM), representing the frequency distribution of terms across documents. To reduce dimensionality and eliminate noise from infrequent terms, sparse term removal was applied using a threshold of 0.99. This procedure produced a refined matrix consisting of 699 documents and 480 unique terms, which served as the input for topic modeling.

Topic modeling was conducted using the Latent Dirichlet Allocation (LDA) algorithm. LDA conceptualizes each document as a probabilistic mixture of latent topics and each topic as a probability distribution over words. The number of topics was predetermined as three to achieve an interpretable and contextually meaningful segmentation aligned with educational policy discourse. Parameter estimation was performed using the Gibbs Sampling algorithm with a fixed random seed to ensure reproducibility and computational stability. In addition to topic modeling, a descriptive spatial analysis was conducted based on available tweet location metadata to examine the geographic distribution of public discourse. Location data were aggregated at the city level to identify regions contributing most prominently to discussions on educational policy. This complementary analysis provided contextual insight into the spatial dynamics of public opinion.

The results of the LDA model were interpreted by examining the probability distribution of terms within each topic and the distribution of topics across documents. High-probability terms were used to derive substantive thematic interpretations. Furthermore, the dominant topic assigned to each tweet was analyzed to assess the overall distribution of public opinion. The findings were visualized using dominant term plots, topic proportion graphs, and geographic distribution charts to enhance interpretability and analytical clarity.

Result and Discussion

Result

Topic modeling using the Latent Dirichlet Allocation (LDA) algorithm was applied to 699 filtered tweets related to educational policy. The model was specified with three topics based on exploratory evaluation to ensure interpretability while maintaining thematic distinctiveness. This approach is consistent with probabilistic topic modeling principles, in which each document is represented as a mixture of latent topics and each topic as a probability distribution over words (Blei et al., 2003). The resulting topic distribution is presented in Figure 1.

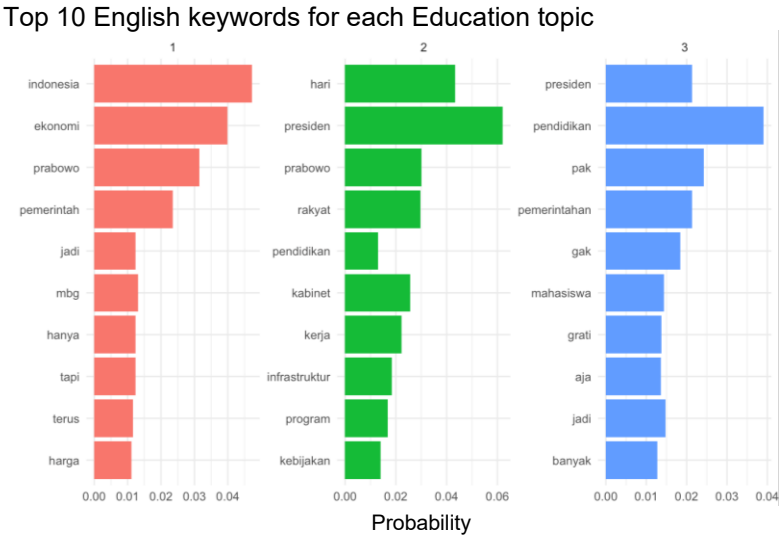


Figure 1. Distribution of Public Opinion Topics on Educational Policy

The distribution demonstrates that Topic 3 accounts for the largest proportion of tweets, followed by Topic 2, while Topic 1 contains the smallest share. The dominance of Topic 3 indicates that educational discourse on Twitter is primarily concentrated on issues directly affecting stakeholders, particularly students. From a probabilistic modeling perspective, this suggests that a substantial proportion of documents in the corpus exhibit higher posterior probabilities associated with Topic 3, indicating thematic salience within the dataset. Similar patterns of dominant thematic concentration have been observed in prior social media–based topic modeling studies (Mujahid et al., 2021; Sun & Yan, 2023).

To examine the semantic structure of each topic, the ten highest-probability words per topic were extracted, as shown in Figure 2.

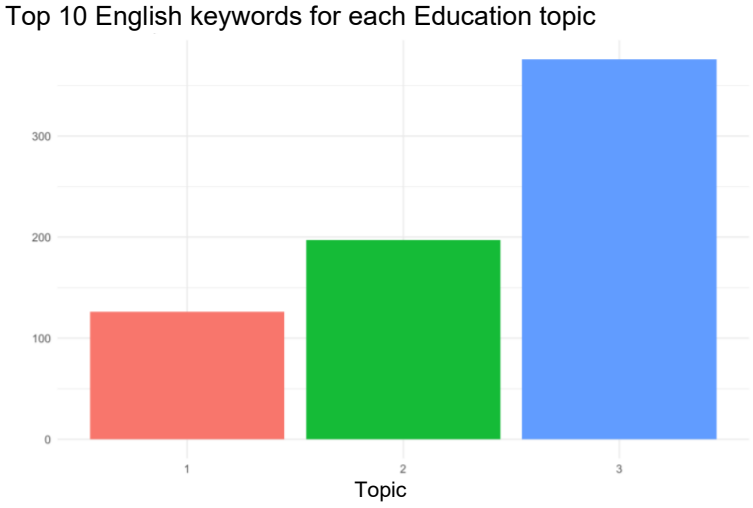


Figure 2. Top Ten Highest-Probability Words per Topic

The word distributions represent β (beta) values derived from the LDA model, reflecting the conditional probability of a word given a topic (Blei et al., 2003). Topic 1 is characterized by macro-level terms such as *Indonesia, economy, government,* and *price*, indicating that educational discussions are frequently embedded within broader national economic and policy narratives. This suggests that a segment of public opinion frames

education as part of macroeconomic governance and national development agendas rather than as an isolated sectoral issue.

Topic 2 is dominated by governance-related terms including *president, public, cabinet, program, policy,* and *infrastructure*. The lexical concentration around political leadership and governmental programs suggests that education is discussed in relation to administrative performance and policy legitimacy. This finding aligns with prior research indicating that public discourse on social media frequently evaluates educational initiatives within broader political accountability frameworks (Zulqarnain & Cahyo, 2023; Negara & Triadi, 2022). Topic 3, which exhibits the highest tweet frequency, contains strongly education-specific terms such as *education, students, government,* and *free*, alongside informal expressions. The co-occurrence of institutional terms and informal language indicates that discourse in this cluster is grounded in direct public experience and personal evaluation. The probabilistic weight of student-related terminology within this topic confirms that the most salient discourse dimension is experiential and stakeholder-centered. This result is consistent with findings that social media platforms function as participatory spaces where affected groups articulate policy-related concerns (Mujahid et al., 2021).

In addition to thematic modeling, a descriptive geographic analysis was conducted to identify the spatial concentration of discourse. The distribution of tweets by city is presented in Figure 3.

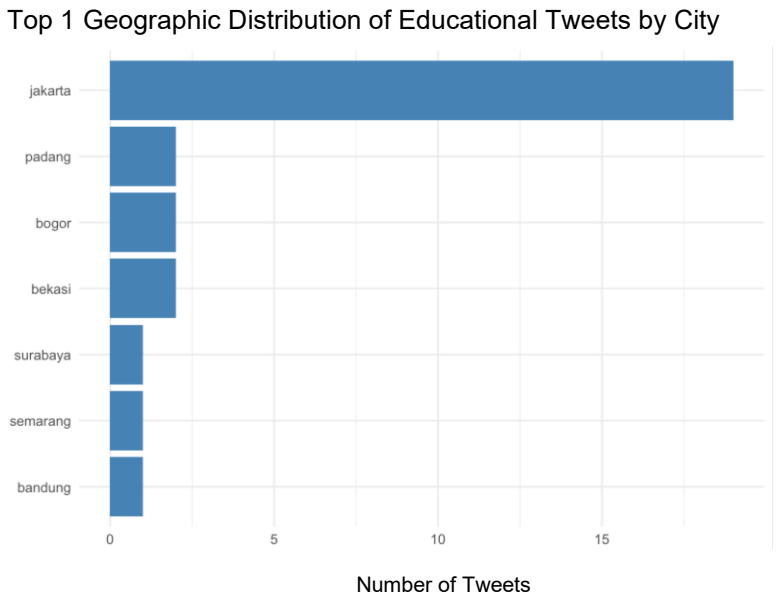


Figure 3. Geographic Distribution of Educational Policy Tweets by City

The results show that Jakarta contributes the highest number of tweets by a substantial margin compared to Padang, Bogor, Bekasi, Surabaya, Semarang, and Bandung. The concentration of discourse in Jakarta suggests a spatial centralization of educational policy discussions. From a digital communication perspective, metropolitan regions particularly political and administrative capitals tend to generate higher levels of policy-related engagement due to increased media exposure, institutional presence, and digital infrastructure. Similar urban-centered discourse patterns have been observed in prior topic modeling studies on Twitter (Zulqarnain & Cahyo, 2023).

The geographic imbalance observed in Figure 3 indicates that educational policy discourse is not evenly distributed across regions. Instead, it is concentrated in urban centers, particularly Jakarta. This spatial pattern complements the thematic findings by demonstrating

that dominant discourse clusters are also geographically centralized. Taken together, the results confirm that public opinion on educational policy in Indonesia is segmented along two dimensions: thematic structure and spatial concentration. The LDA model effectively captured latent thematic patterns consistent with probabilistic topic modeling theory (Blei et al., 2003), while the geographic analysis revealed uneven regional participation in digital discourse. These combined findings provide a structurally grounded empirical representation of how educational policy is discussed within Indonesian Twitter networks.

Discussion

The findings reveal that public discourse on educational policy in Indonesia is structured along distinct thematic and spatial dimensions. The dominance of Topic 3, characterized by student-centered and experiential vocabulary, suggests that educational discussions on Twitter are primarily driven by stakeholders who are directly affected by policy implementation. From a theoretical standpoint, this reflects the participatory nature of social media, where individuals articulate personal experiences and policy evaluations in real time. Previous studies have similarly observed that Twitter functions as a digital public sphere in which affected communities actively engage in policy-related discourse (Mujahid et al., 2021; Sun & Yan, 2023). The prominence of experiential and informal language further indicates that discourse in this topic cluster is grassroots-oriented rather than institutionally framed.

In contrast, Topic 1 situates educational policy within macroeconomic and national governance narratives. The co-occurrence of terms related to economic conditions and government authority suggests that education is frequently framed as part of broader development and political agendas. This supports the argument that sectoral policies are often interpreted by the public through the lens of national economic stability and governance performance. Such framing patterns align with communication theories suggesting that policy issues on social media are rarely isolated; instead, they are embedded within larger socio-political narratives (Hudha Nasution & Prasetya Adhi, 2021). The presence of macro-level discourse demonstrates that public engagement with educational policy extends beyond immediate institutional concerns to encompass systemic governance considerations.

Topic 2 further reinforces the political dimension of educational discourse. The lexical dominance of leadership- and governance-related terms indicates that public conversations frequently evaluate education in relation to executive authority, cabinet performance, and program implementation. This finding highlights the accountability function of social media platforms, where users assess policy effectiveness and governmental legitimacy. Consistent with prior research, topic modeling on Twitter often reveals that discussions surrounding public services are intertwined with political leadership evaluation and administrative credibility (Zulqarnain & Cahyo, 2023; Negara & Triadi, 2022). Thus, educational policy discourse appears to operate at the intersection of sectoral policy debate and political performance assessment.

Beyond thematic segmentation, the geographic concentration of tweets in Jakarta reveals an additional structural characteristic of digital discourse. As Indonesia's political and administrative center, Jakarta naturally serves as a focal point for policy-related engagement. However, the disproportionate contribution from the capital also suggests the presence of regional disparities in digital participation. Urban centers typically exhibit higher levels of internet penetration, media access, and political engagement, which may explain the observed spatial imbalance. This pattern corresponds with broader findings in digital communication research indicating that online political discourse tends to be centralized in metropolitan regions with stronger institutional connectivity (Zulqarnain & Cahyo, 2023). The relatively limited representation from other cities may influence the framing and dominance of certain narratives within national discourse.

Integrating thematic and spatial analyses enhances the explanatory depth of the study. While LDA successfully uncovers latent thematic structures consistent with probabilistic topic modeling theory (Blei et al., 2003), the geographic analysis contextualizes these themes within patterns of regional participation. The convergence of thematic dominance (student-centered discourse) and spatial concentration (urban-centered participation) suggests that educational policy debates on Twitter are shaped by both stakeholder proximity and geographic centralization.

Overall, the findings contribute to the growing body of literature on computational social science and policy analytics by demonstrating how topic modeling can reveal multi-dimensional patterns in public opinion. The segmentation into macro-policy discourse, governance-centered discourse, and experiential stakeholder discourse underscores the complexity of public engagement with educational policy in digital environments. These insights indicate that policymakers should consider not only thematic concerns expressed online but also the geographic concentration of digital voices when interpreting social media–based public opinion data.

Conclusion

This study successfully identified and mapped the segmentation of public opinion regarding educational policy in Indonesia using a Latent Dirichlet Allocation (LDA) approach applied to Twitter data. The modeling results revealed three principal thematic clusters representing distinct dimensions of public discourse. The first topic situates education within broader national policy and economic contexts, the second associates education with political leadership and governmental performance, and the third reflects direct experiential perspectives, particularly from students regarding the implementation and impact of educational policies.

The topic distribution indicates that stakeholder-centered discourse constitutes the largest proportion of public discussion, highlighting strong societal attention toward issues that directly affect everyday educational experiences, such as access to education, tuition costs, and governmental responsibility. The semantic visualization of high-probability keywords further substantiates the interpretative validity of each thematic cluster, demonstrating that educational discourse on social media is inherently intertwined with broader political and governance dynamics. These findings confirm the effectiveness of LDA-based topic modeling in uncovering latent structures within unstructured textual data and in providing a systematic representation of public opinion patterns. By integrating probabilistic modeling with descriptive spatial analysis, the study offers a multidimensional understanding of how educational policy is discussed in digital environments.

From a practical perspective, the results provide empirical insights that may support policymakers in evaluating public concerns and designing more responsive educational policies. Understanding the dominant thematic concerns and the concentration of discourse can assist decision-makers in aligning policy communication strategies with stakeholder expectations. Future research may extend this approach by incorporating temporal analysis to examine changes in discourse over time or sentiment analysis to assess the polarity of public opinion. Such extensions would further enrich the analytical depth and contribute to a more dynamic understanding of educational policy discourse in digital spaces.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no competing interests.

Data Availability

The datasets generated during and analysed during the current study are not publicly available due ethics related to protecting the privacy and confidentiality of research participants but are available from the corresponding author on reasonable request.

References

Arianto, & Anuraga. (2025). Topic modeling analysis of Indonesian policy discourse. *Journal of Data Science and Analytics*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology, 7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data, 2*, Article 5. https://doi.org/10.1186/s40537-015-0015-2

Göçen, A., Ibrahim, M. M., & Khan, A. U. I. (2024). Public attitudes toward higher education using sentiment analysis and topic modeling. *Discover Artificial Intelligence, 4*, Article 83. https://doi.org/10.1007/s44163-024-00195-4

Hu, Y., John, A., Wang, F., Seligmann, D. D., & Kambhampati, S. (2012). ET-LDA: Joint topic modeling for analyzing Twitter feeds. *ACM Transactions on Intelligent Systems and Technology, 4*(2). https://doi.org/10.1145/2661829.2662005

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics, 13*(1), 72–91. https://doi.org/10.1080/19331681.2015.1132401

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Applied Sciences, 11*(18), 8438. https://doi.org/10.3390/app11188438

Negara, E. S., & Triadi, D. (2022). Topic modeling using LDA on Twitter data with Indonesia keyword. *Bulletin of Social Informatics Theory and Application, 5*(2). https://doi.org/10.31763/businta.v5i2.455

Patmawati, & Yusuf. (2021). Topic modeling analysis on Indonesian public policy tweets. *Journal of Data and Information Science*.

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Sentiment analysis and opinion mining on educational data. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10260-y

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.

Sokolova, M., & Bobicev, V. (2016). Topic modeling for Twitter event analysis. *Information Processing & Management, 52*(6). https://doi.org/10.1016/j.ipm.2016.02.005

Sun, J., & Yan, L. (2023). Using topic modeling to understand comments in student evaluations of teaching. *Discover Education, 2*, Article 25. https://doi.org/10.1007/s44217-023-00051-0

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S., & Parnet, O. (2021). A bad workman blames his tweets. *Journal of Communication, 71*(5), 636–661. https://doi.org/10.1093/joc/jqab034

Uthirapathy, S. E., & Sandanam, D. (2023). Topic modelling and opinion analysis on climate change tweets. *Procedia Computer Science, 218*, 1544–1553. https://doi.org/10.1016/j.procs.2023.01.134

Wang, H., et al. (2022). Tweet topics and sentiments relating to distance learning. *Scientific Reports, 12*, Article 12915. https://doi.org/10.1038/s41598-022-12915-w

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery. *Proceedings of EMNLP*. https://doi.org/10.3115/1610075.1610094

Zulqarnain, M. I., & Cahyo, P. W. (2023). Topic modeling and social network analysis on stock discussion tweets. *Indonesian Journal on Data Science*. https://doi.org/10.33096/ijodas.v1i1.15