Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 3, August, Pp. 295-306

Analysis of public opinion on the governor candidate debate using LDA and IndoBERT

Ahmad Abdul Chamid*[1], Ratih Nindyasari[1], Noor Azizah[2], Ahmad Hariyadi[3]
Department of Informatics Engineering, Faculty of Engineering, Universitas Muria Kudus, Indonesia[1]
Department of Information Systems, Faculty of Science and Technology, Universitas Islam Nahdlatul Ulama Jepara, Indonesia[2]
Department of Master of Elementary Education, Faculty of Teacher Training and Education, Universitas Muria Kudus, Indonesia[3]

Article Info
Keywords:
Analysis of Public Opinion, Topic Modeling, Sentiment Classification, Latent Dirichlet Allocation (LDA), BERT

Article history:
Received: January 13, 2025
Accepted: May 24, 2025
Published: August 31, 2025

Cite:
A. A. Chamid, R. Nindyasari, N. Azizah, and A. Hariyadi, "Analysis of Public Opinion on The Governor Candidate Debate Using LDA and IndoBERT", *KINETIK*, vol. 10, no. 3, Aug. 2025.
https://doi.org/10.22219/kinetik.v10i3.2221

*Corresponding author.
Ahmad Abdul Chamid
E-mail address:
abdul.chamid@umk.ac.id

Abstract
The gubernatorial candidate debate was broadcast live streaming through various YouTube channels, which attracted public attention. Many discussions and conversations appeared in the comments section of each YouTube channel that broadcasted the debate. Given the numerous public discussions, it is undoubtedly interesting to analyze the contents of the conversations, as well as the expectations and feedback from the public. However, analyzing conversations in the form of text data will be challenging using conventional methods. Therefore, in this study, public opinion will be analyzed using the topic identification and sentiment classification approaches. Topic identification is conducted to obtain accurate information about what the public is discussing, while sentiment classification is used to determine whether each comment contains positive or negative sentiments. This research is novel because it utilizes data collected from various major media YouTube channels and includes a qualitative analysis of the findings. This study uses public comment data taken from the KPU, NarasiTV, and KompasTV YouTube channels; the results obtained included 4,147 data points. Data preprocessing involves identifying topics using the LDA method, evaluating the LDA model, performing sentiment classification using IndoBERT, and visualizing the results of the public opinion analysis. The results revealed five topics with a perplexity value of -7.7909 and a coherence score of 0.5109. In addition, topic 4 is the most dominant compared to other topics, with 1,146 comments classified as positive sentiment and 504 classified as negative sentiment. Topic 4 reflects how religion, culture, and frequently mentioned figures are perceived and discussed by the public, especially in relation to the gubernatorial election (pilgub) or gubernatorial candidate debates.

1. Introduction

Public conversations are always interesting to analyze, especially during the regional head election (pilkada) season. At the end of 2024, Indonesia will hold simultaneous regional elections, one of which is the gubernatorial election (pilgub) [1]. Several stages must be passed by the candidates, one of which is the gubernatorial candidate debate, which is held publicly and broadcasted via live streaming by several TV stations and the official YouTube channel of the local General Election Commission (KPU). The performance of the candidate pairs during the debate attracted public attention, and many comments were submitted by the public on YouTube. The various comments on the YouTube channel provide a way for the public to convey expressions, hopes, and input to each candidate pair. With the live streaming of the gubernatorial candidate debate, the public can use it to assess the capacity and capabilities of each candidate pair and get to know each candidate better. In addition, the ability to write comments on the YouTube channel is quite helpful for the public in conveying opinions, aspirations, and hopes. However, if one person reads the comments, it will be pretty tiring and time-consuming. Therefore, a special approach is needed to analyze public opinion data to obtain information that can be used to understand what the public wants, whether each candidate pair can be accepted by the public, whether the programs of each candidate pair can convince the public, and what strategies each candidate pair needs to adopt to gain public trust. One approach that can be used to analyze public opinion is Natural Language Processing (NLP) [2], [3], [4], [5], [6].

Several approaches in NLP have been used for public opinion analysis. For example, researchers [7] used the Latent Dirichlet Allocation (LDA) method to identify public discussion topics during the flood disaster that occurred in Jakarta. The data used came from Twitter, utilizing the sentiment analysis approach, which determines public sentiment toward the discussion topics that have been successfully identified. In addition, the sentiment analysis approach has been applied to analyze public opinion regarding the 2020 regional elections using the support vector machine (SVM) method by utilizing data from Twitter, and the results obtained an accuracy value of 87.94% [8]. Previous researchers

also used the SVM method for public sentiment analysis of the 2017 regional elections; the data also came from Twitter, and the results yielded an accuracy value of 91% [9]. The SVM and Bag-of-Words methods were applied in the sentiment analysis process for the 2020 regional elections; the data used also came from Twitter, and the results yielded an accuracy value of 87.5% [10]. Previously, [11] used the naïve Bayes (NB) and SVM methods for sentiment analysis using data sourced from Twitter; the results showed that the NB accuracy value was 92.2%, outperforming the SVM method. In addition to several classical machine learning methods tested for sentiment analysis in regional elections, [12] previously tried using the deep learning method, namely the Convolution Neural Network (CNN). The data used also came from Twitter, and the CNN method was applied by adjusting the parameters during the training process, resulting in an accuracy value of 90%.

Based on previous research, it is known that sentiment analysis has been widely used to analyze public opinion on regional elections. Various methods have also been tried; however, in general, researchers only focus on sentiment classification, which only produces sentences or opinions containing either positive or negative sentiment. As a result, it has not provided information that can answer what the public likes and dislikes. In this study, the LDA method is used to conduct deeper information mining on comments written by the public. The LDA method identifies topics discussed by the public, and the successfully identified topics will be classified by their sentiment using the IndoBERT method. Previously, the LDA method was applied to identify topics in Arabic Twitter data during the COVID-19 pandemic, where it was found that LDA can identify three major topics in public discussions [13]. In addition, the LDA method was also used to identify topics of conversation on Twitter related to the COVID-19 pandemic, resulting in 20 topics discussed in the conversation [14]. Researchers [15] used a topic modeling approach to identify topics in Western (the Euronews) and Eastern (the Kyiv Post) media related to the Russia-Ukraine war during 2022-2023. The method used was LDA, and the results obtained were able to provide a comprehensive analysis of the nuances in the narrative, public perception, and policy in the ongoing war. Researchers [16] applied the LDA method to find topics from reviews posted by hotel guests. A framework was developed to organize the identified themes and describe the aesthetic experience of hotel guests. Researchers [17] used a topic modeling approach with the LDA method to analyze topics and opinions discussed by guests regarding climate change by utilizing Twitter data. Lastly, researchers [18] used the LDA method to identify topics in conversations among Malaysian citizens on Twitter regarding the government's response to the COVID-19 disaster.

Based on what has been described above, sentiment analysis at the sentence level cannot provide in-depth information. In addition, by only identifying topics that arise from various opinions, we cannot gain insights into what the public likes and dislikes. Therefore, in this study, the LDA and IndoBERT methods are proposed for public opinion analysis. The LDA method is used to identify topics that frequently appear in public conversations, while the IndoBERT method is used to classify sentiment. Detailed and in-depth information will be obtained by identifying topics and categorizing sentiments in each opinion sentence. The difference between this research and previous research lies in the data sources obtained and the analysis results presented in qualitative detail.

2. Research Method

This research was conducted using the Python programming language in Google Colab. The research stage process, as shown in Figure 1, is explained as follows:

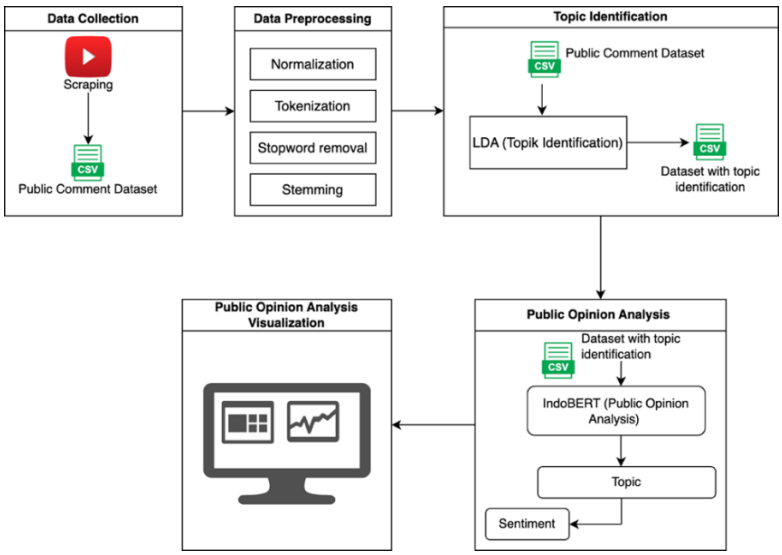

Figure 1. Research Method

2.1 Data Collection

This study utilizes public comment data taken from the YouTube platform, specifically data from the first and second debates of the Candidates for Governor (Cagub) and Candidates for Deputy Governor (Cawagub) of Central Java (Jateng) 2024 from the KPU, KompasTV, and NarasiTV YouTube Channels. The data retrieval process employs the scraping technique using Python programming and Google Colab. During the scraping process, the YouTube API, which is available for free, was utilized, and the scraping process took place from November 18 to19, 2024. The scraping results obtained 4,147 public comments, some examples of which are shown in Table 1.

Table 1. The Example of Public Comments

Index	Public Comments
1	Itu cuman ilusi ..tak akan ngaruh sama sekali buat masyarakat..ngaruh yang dekat sama paslonnya doank yakiiiinnn…
2	Pilih 02 karena ngalap berkahnya kyai… (Saya)
3	Terlalu muluk² programnya 02
4	Daya sains harus di majukan.
5	Jawa tengah ,megah dan Indah semoga berjaya, semua nya ❤❤❤❤

2.2 Data Preprocessing

The data obtained is in an unstructured form, so data preprocessing is required. The stages of data preprocessing include normalization, which is the process of standardizing text to ensure data consistency and reduce variations in words that have similar meanings. The stages of normalization include case folding, which changes all letters to lowercase to ensure consistency, removing punctuation and symbols by eliminating irrelevant characters such as punctuation, numbers, and special symbols, and converting slang or non-standard language into standard forms [19], [20].

Tokenization is the process of breaking text into smaller units called tokens, usually in the form of words or phrases. The aim is to facilitate analysis by identifying word or sentence boundaries [21]. Stopword removal is used to eliminate common words that do not have significant meaning in the analysis, such as " which", "and", "in", and so on. The goal is to reduce noise in the data and focus on more meaningful words [22], [23]. Stemming involves changing words to their essential form by removing affixes such as prefixes and suffixes; specifically in Indonesian, libraries such as Sastrawi are often used to perform stemming [24]. The results of data preprocessing are shown in Table 2.

Table 2. The Result of Data Preprocessing

Index	Public Comments
1	cuman ilusi ngaruh masyarakat ngaruh paslonnya doank yakiiiinnn (*It's just an illusion that influences the public and only influences the candidate, I'm sure*)
2	pilih 02 ngalap berkahnya kyai (*choose 02 seek blessings kyai*)
3	muluk² programnya 02 (*02's program is grandiose 02*)
4	daya sains majukan (*advancing scientific power*)
5	jawa megah indah semoga Berjaya (*Java is magnificent and beautiful, good luck*)

2.3 Topic Identification

Topic identification using the Latent Dirichlet Allocation (LDA) method was chosen because it has been proven effective in identifying topics in conversational sentences [15], [25], [26]. The topic identification process begins by determining how many topics will be identified or by applying an approach to assess how precisely the topic can be identified by evaluating each identification result. A list of frequently occurring words will be checked for each topic; generally, the top 10-15 words will reflect the topic. Next, the performance of the LDA model will be evaluated to measure how well the model can predict the dataset using the perplexity matrix [27], [28]. In the context of topic modeling, perplexity helps determine how well the model can explain the distribution of words in a document. In addition, a coherence score is used to reflect the topic's quality because it measures how often words in one topic appear together [29]. Next, the topic identification results will be visualized to facilitate the analysis process, and the topic identification results will be saved in CSV format.

2.4 Public Opinion Analysis

After successfully identifying the topic, the next step is to analyze public opinion. The approach involves classifying each opinion sentence as either positive or negative. The classification process uses IndoBert, a version of BERT trained on a large text corpus in Indonesian, making it more effective for NLP tasks in the language, such as sentiment analysis, information extraction, translation, and text classification [30], [31], [32]. This allows for the identification of both the topic and sentiment within a single opinion sentence, with the results saved in CSV format.

2.5 Visualization

Once the topics have been successfully identified and sentiments classified in each opinion sentence, the results will be visualized to facilitate the analysis process and interpreted into information that is easy for the public to understand. The visualization process uses various diagrams tailored to the needs of the analysis, ensuring clarity. One of the diagrams used to visualize the results of topic identification is Python LDA Visualization (pyLDAvis), an interactive visualization library that helps explore and understand LDA results. pyLDAvis allows users to visualize topics generated by the LDA model, explore relationships between topics intuitively, and view the dominant words in each topic, enhancing the understanding of their contribution.

3. Results and Discussion

The topic identification process is carried out by building an LDA model. The model is then evaluated to obtain the appropriate number of topics using the perplexity matrix and coherence score. This evaluation process is implemented through trials in the range of 2 to 10 topics, with the results shown in Figure 2.

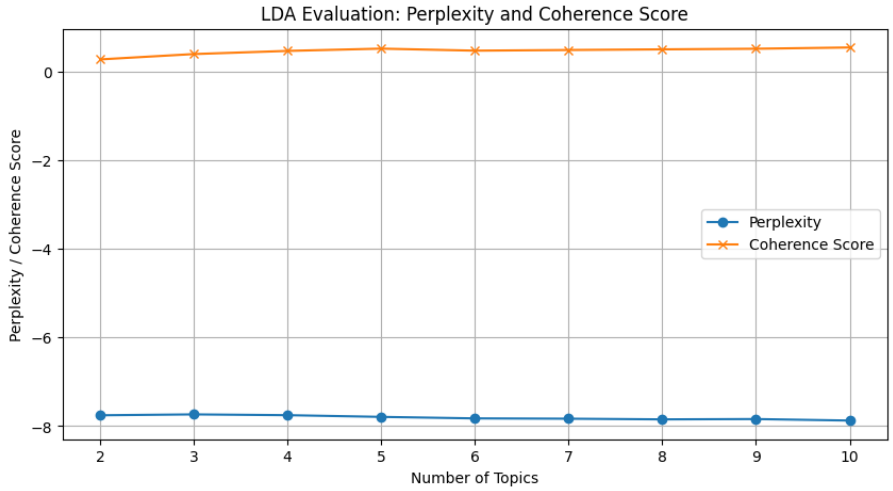

Figure 2. The Result of the LDA Evaluation

Based on Figure 2, the process of determining the number of topics is defined within the range of 2 to 10. The results show that identifiying 10 topics yields a perplexity value of -7.8738 and a coherence score of 0.5368. However, with the increase in identified topics, there is a concern that the analysis of public conversations may not focus on the core of the discussion. Furthermore, observations were made on five topics, which showed a perplexity value of -7.7909 and a coherence score of 0.5109. Therefore, this study determined that topic identification would utilize five topics, considering the perplexity and coherence scores, which remain quite good. Additionally, after five topics, coherence tends to fluctuate and generally increases. The lower (more negative) the perplexity value, the better. Conversely, the higher the coherence value, the better (closer to 1) [27], [29]. The perplexity value is not always the sole indicator; sometimes, even with a low perplexity, the results may be less interpretable. Therefore, it is essential to combine perplexity with the coherence score. Perplexity does not always directly correlate with topic interpretability. Coherence better reflects the quality of a topic because it measures how often words in a topic appear together [27], [29]. Thus, high coherence is crucial to ensuring that the resulting topics are relevant and easy to understand. In addition, the fewer topics there are, the more specific the resulting analysis will be. In this study, five topics were selected based on the results of perplexity and coherence scores.

Next, visualization is carried out using five topics, sequentially visualizing the results of identifying topics 1 to 5, as seen in Figure 3. The visualization results for Topic 1 appear to have the largest diameter compared to the other topics, indicating that Topic 1 is frequently discussed by the public. The five most relevant terms for Topic 1 are "andika", "jateng", "polisi", "hendi", and "tni". The analysis results for Topic 1 show that the public often uses the term "andika",

which represents the name of one of the gubernatorial candidates. If seen in more detail, the following term is "polisi," which is the agency of origin for one of the gubernatorial candidates.

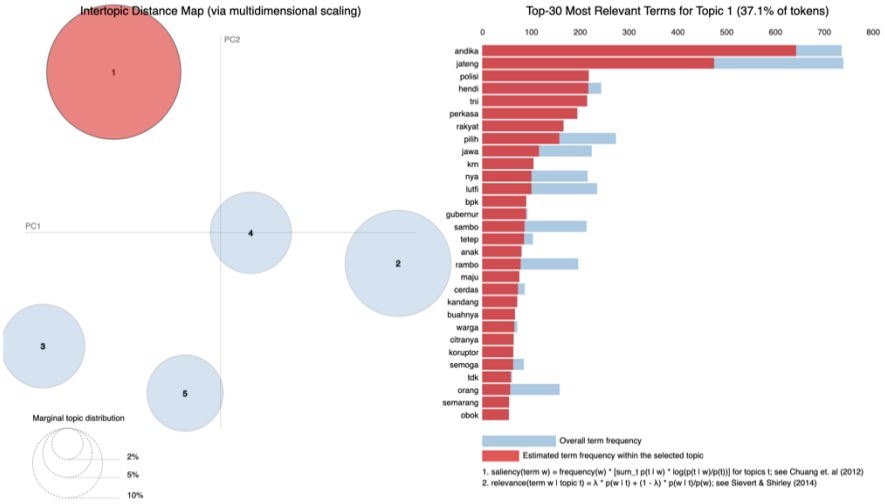

Figure 3. The Visualization of Topic-1

The visualization of Topic 2, as shown in Figure 4, contains 23.1% of the tokens. In addition, the frequency of terms that often appear, along with the five most relevant terms for Topic 2, includes "kalah", "dpat", "anti", "santri", and "blum". The term "kalah" is the most relevant term in Topic 2, representing support or psywar from supporters of each candidate pair. Upon closer observation, the term "santri" refers to the deputy governor candidate from one of the candidate pairs. Notably, one of the candidates comes from Islamic boarding schools, commonly referred to as “santri” in Central Java Province (Jateng). The cultural culture in Central Java is very familiar with the term "santri", and it is likely that the term is not only in direct conversations in public places but also in discussion on online platforms.

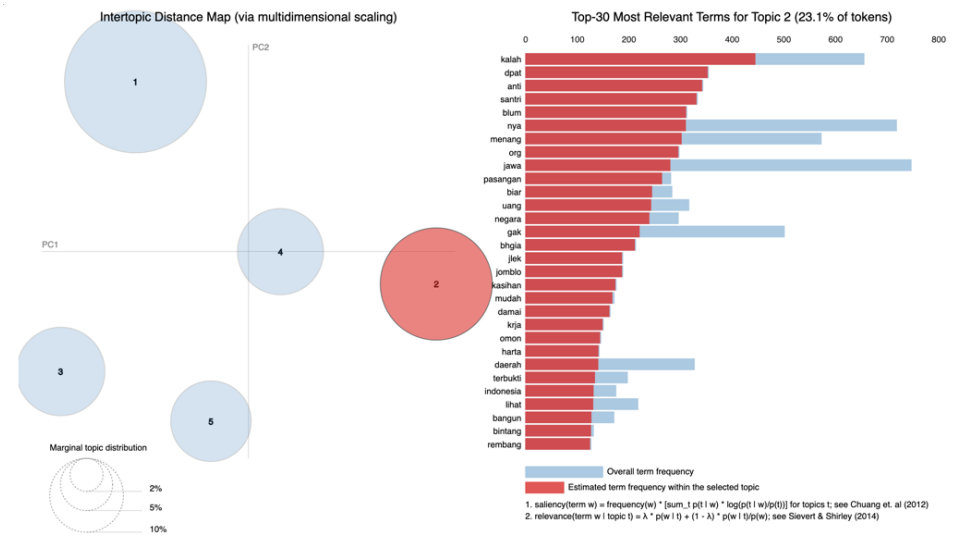

Figure 4. The Visualization of Topic-2

Next, Topic 3 is visualized, as shown in Figure 5. The result is that 14.3% of tokens can be represented in Topic 3. The estimated frequency of terms often appearing in Topic 3 is visualized in red, while the overall frequency is visualized in light blue. There are top 30 terms that are relevant to Topic 3, with the top five including "banteng", "all", "pilih", "andika", and "wes". It can be seen that "banteng" is the top term in Topic 3, representing one of the political parties (parpol) in Indonesia. The term "banteng" is often used by supporters or the general public to refer to this political party. Additionally, the term "banteng" sounds like the name of an animal, and one of the parties supporting one of the candidate pairs indeed has a bull symbol or logo. Thus, it is clear that the term "banteng" represents the supporting party of one of the candidate pairs. Furthermore, the term “andika” represents the name of one of the candidate pairs

and is related to the term “banteng”. In fact, the party that is identical to the term “banteng” supports the candidate pair associated with the term “andika”.

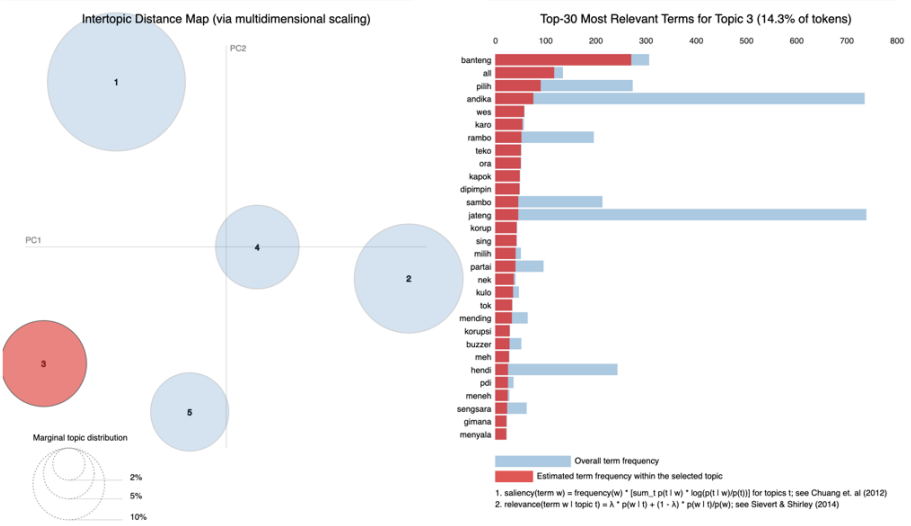

Figure 5. The Visualization of Topic-3

Figure 6 is a visualization of Topic 4; it is known that there are 13.6% of tokens, and the top 30 relevant terms in Topic 4 are displayed. Based on the estimated frequency of terms that often appear in Topic 4, the top five terms are "orang", "sambo", "rambo", "jateng", and "luthfi". Upon closer observation, it is known that the term "sambo" represents one of the former police officers who was convicted of murder, and the word "sambo" is also part of the name of this convicted officer. The case has been reported in national and international news. The term "sambo" has become familiar in Indonesian society and is often a topic of conversation in public places and online platforms. Whenever a police officer makes a mistake, the public will always associate it with the Sambo incident. Interestingly, the term "luthfi," which represents one of the candidate pairs, has a background in the police, and it appears that the term"luthfi" is related to the term "sambo". Therefore, the identification results of Topic 4 are related to one of the candidates with a police background, and the community has high hopes that the candidate pair if elected, will not act like Sambo. In addition, it reminds the community about the background and track record of one of the candidates and fosters hope that leadership like Sambo will not happen again.

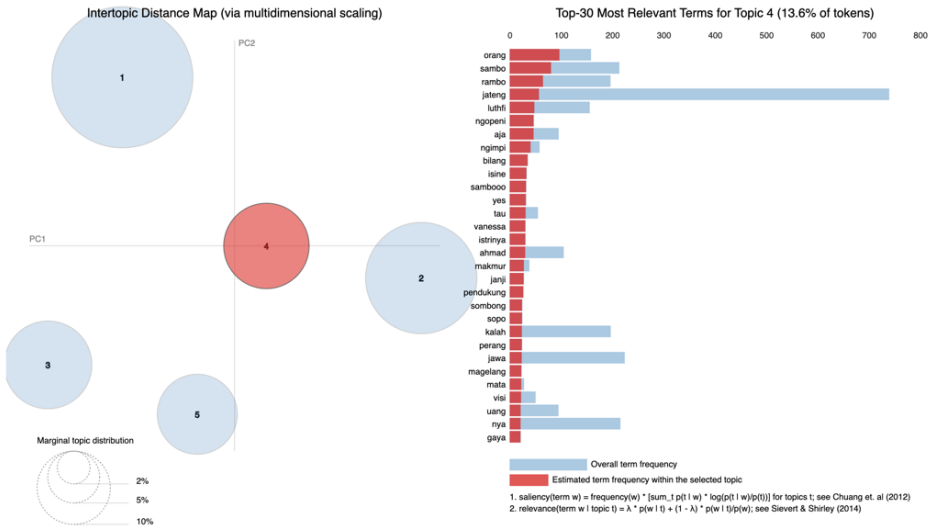

Figure 6. The Visualization of Topic-4

The visualization results of Topic 5, as shown in Figure 7, contain 11.9% of the tokens and display the top 30 terms relevant to Topic 5. The estimated frequency of terms that often appear in Topic 5 include the top 5: "yasin", "lutfi", "jateng", "gus", and "pilihan". Based on the frequency of terms that often appear in Topic 5, it is evident that

"yasin" and "lutfi" represent one of the candidate pairs. In addition, the term "gus" represents one of the candidate pairs' representatives and is closely related to the term "yasin". The term "gus" refers to Javanese culture, specifically a term for a child of a kyai in the Islamic boarding school environment. One of the representatives of the candidate pairs has a background in Islamic boarding schools and is the child of a highly respected kyai in Central Java. Therefore, Topic 5 is more dominant in containing discussions related to one of the candidate pairs related to the terms "yasin" and "lutfi".

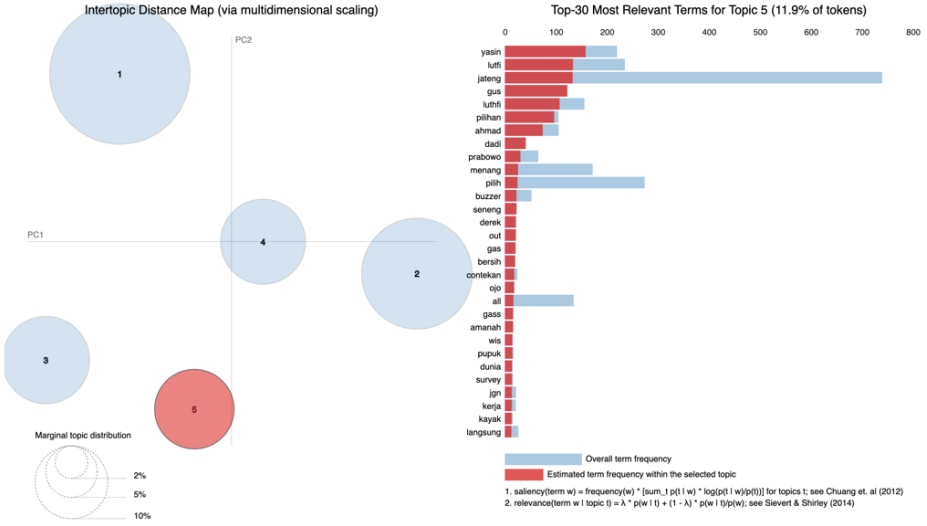

Figure 7. The Visualization of Topic-5

After obtaining the LDA model with 5 topics and visualizing each topic, a more in-depth analysis is conducted to find the dominant words. The results are visualized using WorldCloud. Figure 8 illustrates the dominant words in Topic 1; it is evident that the top 5 dominant words consist of "teko," "butuh," "jam," "debate," and "terbukti." The results show that campaign-motivated words dominate Topic 1, as each candidate's supporters are trying to campaign the candidate they support. In addition, the word "jatim" appears among the dominant words, indicating that the condition of Central Java (Jateng) is compared to East Java (Jatim). The community tries to provide a comparison with East Java Province, which can be interpreted as one of the hopes of the community, but what the community wants—whether infrastructure, education, or welfare—is not yet or is less clear. Thus, Topic 1 can be interpreted as a "political campaign", where each candidate's supporters flood the comments to advocate for the candidate they support.

Figure 8. Visualization of Dominant Words in Topic 1

The visualization of dominant words in Topic 2 is shown in Figure 9. The results show the top five dominant words, namely "luthfi", "jateng", "kesejahteraan", "bilang", and "tau". In Topic 5, the word "luthfi" appears, which is the name of one of the candidates, while the word "kesejahteraan" expresses the community's hope that each candidate will improve the welfare of the Central Java community if elected. Topic 2 discusses welfare issues in Central Java, with the spotlight on the figure "luthfi", which can be related to his campaign promises or statements made during the debate.

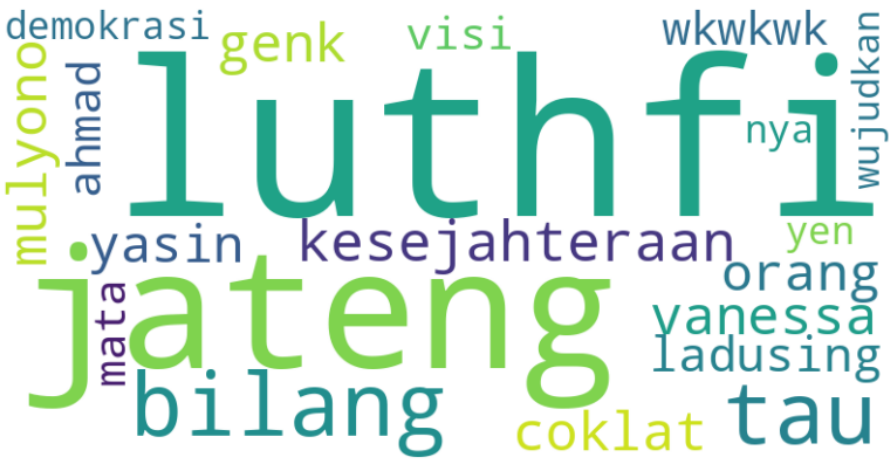

Figure 9. Visualization of Dominant Words in Topic 2

Figure 10 shows the visualization of dominant words in Topic 3; the dominant words include "polisi", "tni", "jateng", "rakyat", and "orang". Topic 3 revolves around the role of the police and TNI in maintaining security and serving the people of Central Java. It can also include discussions about the interaction between security institutions and the people, encompassing criticism, appreciation, or hope. Potential issues can be related to the candidate's vision or policies regarding security in Central Java, and criticism or support for how the police and TNI fulfill their duties is also highlighted. The social significance of the identification of Topic 3 reflects the importance of security and stability in public discussions, especially during the election process or other relevant social issues. The background of each gubernatorial candidate is rooted in the police and TNI institutions.

Figure 10. Visualization of dominant words in Topic 3

The visualization of dominant words in Topic 4, as shown in Figure 11, shows that the top five most dominant words include "yasin", "luthfi", "gus", "santri", and "dpat". It can be seen that Topic 4 focuses on religious figures, the santri community and the pesantren culture, in addition to the discussion of how specific candidates have relationships with religious communities, especially santri and pesantren. Potential issues discussed include support from the santri community for specific candidates, and the role of religious values in local politics, including how santri or pesantren are involved in elections or candidate debates. The words "gus" and "santri" indicate the importance of religious figures and pesantren traditions in political and social discussions in Central Java society. The words "yasin" and "luthfi" can indicate specific figures who influence the community. Topic 4 seems to focus on the role of the santri community, religious figures, and pesantren values in local political discussions in Central Java. These words reflect how the public perceives religion, culture, and figures such as Yasin and Luthfi, especially concerning the gubernatorial election (pilgub) or gubernatorial candidate debates. For example, the comment “Jawa tengah butuh nasionalis relugius ... Gus yasin perwakilan santri” can influence people who view religiosity as very important.

Figure 11. Visualization of Dominant Words in Topic 4

Figure 12 shows the visualization of dominant words in Topic 5. The results show the top five dominant words in Topic 5, namely "sambo", "rambo", "pilih", "milih", and "korup". Topic 5 discusses issues related to corruption, controversial figures, and political choices in the context of elections or local political discussions in Central Java. In addition, Topic 5 seems to discuss political choices in elections, which are influenced by significant issues such as corruption and the reputation of specific figures. Names or terms such as "sambo" and "rambo" refer to figures or symbols relevant to political discussions. Comparisons between political figures use certain nicknames or terms to discredit or highlight the characteristics of each candidate pair. Topic 5 focuses on the issue of candidate integrity, corruption, and people's political decisions. This discussion reflects how specific figures become symbols or centers of attention in public conversations, often accompanied by criticism or satire of the current political situation.

Figure 12. Visualization of Dominant Words in Topic 5

Next, public opinion analysis is conducted using a sentiment analysis approach. The process employs IndoBERT, one of the models explicitly developed using Indonesian, with the central concept based on the BERT model. Additionally, IndoBERT can classify sentiment effectively and can be used for other tasks related to Indonesian text; it can also be combined with other methods [33], [34], [35]. Examples of topic identification and sentiment classification results are shown in Table 3.

Table 1 shows that the dataset, initially comprising public comments during the gubernatorial candidate debate, has been preprocessed, and each comment has been identified for its topic and sentiment. For example, the first comment has been identified as Topic 1 and classified as positive sentiment. In IndoBERT, “LABEL_1” = Positive and “LABEL_0” = Negative [36], [37]. The results of the public opinion analysis are visualized in Figure 13 to provide a comprehensive overview.

Table 3. The Result of The Analysis of Public Opinion

No	Comment	Topic	Sentiment
1	['cuman', 'ilusi', 'ngaruh', 'masyarakat', 'ngaruh', 'paslonnya', 'doank', 'yakiiiinnn']	0	LABEL_1
2	['pilih', '02', 'ngalap', 'berkahnya', 'kyai']	2	LABEL_1
3	['muluk²', 'programnya', '02']	2	LABEL_1
4	['daya', 'sains', 'majukan']	1	LABEL_0
5	['jawa', 'megah', 'indah', 'semoga', 'berjaya']	1	LABEL_0

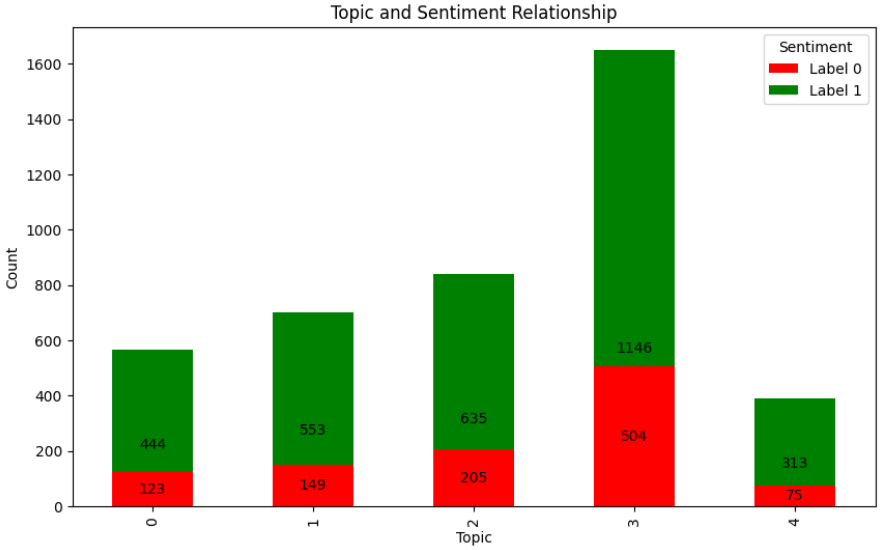

Figure 13. Topic and Sentiment Relationship

Figure 13 shows the relationship between topics and sentiment regarding public comments. Based on the data, it is evident that Topic 4 is very dominant compared to other topics, classified as positive sentiment with 1,146 data and negative sentiment with 504 data. Topic 4 is interpreted as the role of the santri community, religious figures, and pesantren values in local political discussions in Central Java. These words reflect how the public perceives religion, culture, and figures such as Yasin and Luthfi, especially regarding the gubernatorial election (pilgub) or gubernatorial candidate debates. The data on Topic 4 indicates that more public comments are classified as positive sentiment than negative sentiment. All conversations on the online platform reflect a significant number of people discussing Topic 4, presenting an opportunity for one of the candidate pairs related to the topic to gain votes during the election. Based on the data sources obtained, the selected YouTube channels are official media from the KPU and major media outlets known to be professional, independent, and trusted, which alleviates concern about data bias.

4. Conclusion

Public opinion analysis of the gubernatorial candidate debate was conducted using public comments on the online platform (YouTube Channel). The process involved identifying the topics discussed using the LDA method. The LDA model was built and evaluated using perplexity and coherence score matrices. The results obtained a perplexity value of -7.7909 and a coherence score of 0.5109, with 5 Topics identified. Each topic was successfully analyzed in detail, and a classification was carried out for each comment using IndoBERT. Based on the analysis of the relationship between topics and sentiment, the results showed that Topic 4 was the most dominant topic, with 1,600 comments compared to other topics. Topic 4 reflects how the public perceives religion, culture, and figures such as Yasin and Luthfi, especially concerning the gubernatorial election (pilgub) or the gubernatorial candidate debate. In addition, in Topic 4, more positive sentiment was classified than negative, indicating an opportunity for the candidate pair who are often associated with it. Further research can be conducted to determine whether there is a relationship between this study's results and the gubernatorial election outcomes. Meanwhile, a more comprehensive sentiment classification model can be developed using the latest deep learning methods.

Acknowledgment

We would like to thank LPPM Universitas Muria Kudus for supporting the research with the “Skema Kerjasama Nasional” and our colleagues who have provided support until publication.

References

[1] Komisi Pemilihan Umum, *Peraturan Komisi Pemilihan Umum*. Indonesia, 2024.

[2] I. Gjorshoska, A. Dedinec, J. Prodanova, A. Dedinec, and L. Kocarev, "Public perception of waste regulations implementation. Natural language processing vs real GHG emission reduction modeling," *Ecol Inform*, vol. 76, Sep. 2023. https://doi.org/10.1016/j.ecoinf.2023.102130

[3] O. Olabanjo *et al.*, "From Twitter to Aso-Rock: A sentiment analysis framework for understanding Nigeria 2023 presidential election," *Heliyon*, vol. 9, no. 5, May 2023. https://doi.org/10.1016/j.heliyon.2023.e16085

[4] S. Ha and E. Grubert, "Hybridizing qualitative coding with natural language processing and deep learning to assess public comments: A case study of the clean power plan," *Energy Res Soc Sci*, vol. 98, Apr. 2023. https://doi.org/10.1016/j.erss.2023.103016

[5] A. A. Chamid, Widowati, and R. Kusumaningrum, "Multi-Label Text Classification on Indonesian User Reviews Using Semi-Supervised Graph Neural Networks," *ICIC Express Letters*, vol. 17, no. 10, pp. 1075–1084, 2023. https://doi.org/10.24507/icicel.17.10.1075

[6] A. A. Chamid, Widowati, and R. Kusumaningrum, "Graph-Based Semi-Supervised Deep Learning for Indonesian Aspect-Based Sentiment Analysis," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 5, 2023. https://doi.org/10.3390/bdcc7010005

[7] M. C. Rahmadan, A. N. Hidayanto, D. S. Ekasari, B. Purwandari, and Theresiawati, "Sentiment Analysis and Topic Modelling Using the LDA Method related to the Flood Disaster in Jakarta on Twitter," in *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) Sentiment*, 2020, pp. 126–130. https://doi.org/10.1109/ICIMCIS51567.2020.9354320

[8] M. Paramarta and J. B. B. Darmawan, "Implementasi Metode Support Vector Machine dalam Analisis Sentimen Opini Masyarakat Terhadap Pilkada 2020 pada Media Sosial Twitter," in *Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XVIII*, Nov. 2023, pp. 836–841.

[9] A. Rahmawati, A. Marjuni, and J. Zeniarja, "Analisis Sentimen Publik Pada Media Sosial Twitter Terhadap Pelaksanaan Pilkada Serentak Menggunakan Algoritma Support Vector Machine," *CCIT Journal*, vol. 10, no. 2, pp. 197–206, 2017. https://doi.org/10.33050/ccit.v10i2.539

[10] R. Pohan *et al.*, "Implementasi Algoritma Support Vector Machine dan Model Bag-of-Words dalam Analisis Sentimen mengenai PILKADA 2020 pada Pengguna Twitter," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 10, pp. 4924–4931, 2022.

[11] A. Muzaki and A. Witanti, "Sentimen Analisis Masyarakat Di Twitter Terhadap Pilkada 2020 Ditengah Pandemic Covid-19 Dengan Metode NaïVe Bayes Classifier," *Jurnal Teknik Informatika (Jutif)*, vol. 2, no. 2, pp. 101–107, 2021. https://doi.org/10.20884/1.jutif.2021.2.2.51

[12] S. N. Listyarini and D. A. Anggoro, "Analisis Sentimen Pilkada di Tengah Pandemi Covid-19 Menggunakan Convolution Neural Network (CNN)," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 1, no. 7, pp. 261–268, 2021. https://doi.org/10.52436/1.jpti.60

[13] N. Habbat, H. Anoun, and L. Hassouni, "Sentiment Analysis and Topic Modeling on Arabic Twitter Data during Covid-19 Pandemic," *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*, vol. 2, no. 1, pp. 60–67, 2022. https://doi.org/10.47540/ijias.v2i1.432

[14] I. Alagha, "Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives," *Journal of Information Science Theory and Practice*, vol. 9, no. 1, pp. 35–53, 2021. https://doi.org/10.1633/JISTaP.2021.9.1.3

[15] A. Verbytska, "Topic modelling as a method for framing analysis of news coverage of the Russia-Ukraine war in 2022–2023," *Lang Commun*, vol. 99, pp. 174–193, Nov. 2024. https://doi.org/10.1016/j.langcom.2024.10.004

[16] S. Ying, "Guests' Aesthetic experience with lifestyle hotels: An application of LDA topic modelling analysis," *Heliyon*, vol. 10, no. 16, Aug. 2024. https://doi.org/10.1016/j.heliyon.2024.e35894

[17] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA and BERT Model.," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 908–917. https://doi.org/10.1016/j.procs.2023.01.071

[18] M. N. P. Ma'ady, A. F. A. Rahim, T. S. N. Syahda, A. F. Rizqi, and M. C. A. Ratna, "Malaysia Citizen Sentiment on Government Response Towards Covid-19 Disaster Management: Using LDA-based Topic Visualization on Twitter," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 561–569. https://doi.org/10.1016/j.procs.2024.03.040

[19] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," Nov. 01, 2024, *Elsevier Ireland Ltd*. https://doi.org/10.1016/j.cosrev.2024.100664

[20] A. A. Chamid, Widowati, and R. Kusumaningrum, "Labeling Consistency Test of Multi-Label Data for Aspect and Sentiment Classification Using the Cohen Kappa Method," *Ingénierie des Systèmes d'Information*, vol. 29, no. 1, pp. 161–167, 2024. https://doi.org/10.18280/isi.290118

[21] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," *Telematics and Informatics Reports*, vol. 16, Dec. 2024. https://doi.org/10.1016/j.teler.2024.100173

[22] A. A. Firdaus, A. Yudhana, and I. Riadi, "Public Opinion Analysis of Presidential Candidate Using Naïve Bayes Method," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, May 2023. https://doi.org/10.22219/kinetik.v8i2.1686

[23] A. A. Chamid, W. Widowati, and R. Kusumaningrum, "Text data labeling process for semi-supervised learning modeling," *12TH INTERNATIONAL SEMINAR ON NEW PARADIGM AND INNOVATION ON NATURAL SCIENCES AND ITS APPLICATIONS (12TH ISNPINSA): Contribution of Science and Technology in the Changing World*, vol. 3165, p. 030011, 2024. https://doi.org/10.1063/5.0216320

[24] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 375–380, Oct. 2019. https://doi.org/10.22219/kinetik.v4i4.912

[25] J. Zimmermann, L. E. Champagne, J. M. Dickens, and B. T. Hazen, "Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling," *Decis Support Syst*, vol. 185, Oct. 2024. https://doi.org/10.1016/j.dss.2024.114310

[26] Y. Jiang, M. Fu, J. Fang, M. Rossi, Y. Wang, and C. W. Tan, "Advancing an LDA-GMM-CorEx topic model with prior domain knowledge in information systems research," *Information and Management*, vol. 62, no. 2, Mar. 2025. https://doi.org/10.1016/j.im.2024.104097

[27] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and D. P. Radicioni, "Semantic coherence markers: The contribution of perplexity metrics," *Artif Intell Med*, vol. 134, Dec. 2022. https://doi.org/10.1016/j.artmed.2022.102393

[28] R. He, C. Palominos, H. Zhang, M. F. Alonso-Sánchez, L. Palaniyappan, and W. Hinzen, "Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models," *Psychiatry Res*, vol. 333, Mar. 2024. https://doi.org/10.1016/j.psychres.2024.115752

[29] T. Cohen, W. Xu, Y. Guo, S. Pakhomov, and G. Leroy, "Coherence and comprehensibility: Large language models predict lay understanding of health-related content," *J Biomed Inform*, vol. 161, Jan. 2025. https://doi.org/10.1016/j.jbi.2024.104758

[30] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *J Informetr*, vol. 14, no. 3, Aug. 2020. https://doi.org/10.1016/j.joi.2020.101055

[31] J. Liu, R. Long, H. Chen, M. Wu, W. Ma, and Q. Li, "Topic-sentiment analysis of citizen environmental complaints in China: Using a Stacking-BERT model," *J Environ Manage*, vol. 371, Dec. 2024, doi: 10.1016/j.jenvman.2024.123112.

[32] J. Lim and J. Hwang, "Exploring diverse interests of collaborators in smart cities: A topic analysis using LDA and BERT," *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e30367.

[33] H. J. Juandri, H. Hasmawati, and B. Bunyamin, "Aspect-Level Sentiment Analysis on GoPay App Reviews Using Multilayer Perceptron and Word Embeddings," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Aug. 2024. https://doi.org/10.22219/kinetik.v9i4.2041

[34] R. A. Rajagede, "Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 11–18, Feb. 2021. https://doi.org/10.22219/kinetik.v6i1.1196

[35] A. Salsabil, E. B. Setiawan, and I. Kurniawan, "Content-based filtering movie recommender system using semantic approach with recurrent neural network classification and SGD," *Computer Network, Computing, Electronics, and Control Journal*, vol. 9, no. 2, pp. 193–202, 2024. https://doi.org/10.22219/kinetik.v9i2.1940

[36] A. B. Y. A. Putra, Y. Sibaroni, and A. F. Ihsan, "Disinformation Detection on 2024 Indonesia Presidential Election using IndoBERT," in *2023 International Conference on Data Science and Its Applications, ICoDSA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 350–355. https://doi.org/10.1109/ICoDSA58501.2023.10277572

[37] R. I. Yulfa, B. H. Setiawan, G. G. Lourensius, and K. Purwandari, "Enhancing Hate Speech Detection in Social Media Using IndoBERT Model: A Study of Sentiment Analysis during the 2024 Indonesia Presidential Election," in *ICCA 2023 - 2023 5th International Conference on Computer and Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. https://doi.org/10.1109/ICCA59364.2023.10401700