*Edueksos: Jurnal Pendidikan Sosial dan Ekonomi*

p-ISSN: 2252-9942 e-ISSN: 2548-5008
https://syekhnurjati.ac.id/jurnal/index.php/edueksos

Volume XV, Number 01, May 2026

# Deliberation or Aggression? Unpacking Toxic Disinhibition on Tiktok's Political and Educational Discourse

**Roul Alvaro Prasetyo[1], Elsa Nurhalisa[2], Budi Mulyono[3]**
Universitas Negeri Yogyakarta, Indonesia[1,2,3]
roulalvaro.2025@student.uny.ac.id. ; elsanurhalisa.2025@student.uny.ac.id. ; budi.mulyono@uny.ac.id.

***Article History***
*Received:*
*19-04-2026*
*Revised:*
*14-05-2026*
*Accepted:*
*18-05-2026*
*Available online:*
*30-05-2026*

**ABSTRACT**

This study aims to analyze the forms of citizens' expressions and participation in the digital public sphere, as well as to uncover public opinion anomalies on a state official's policy post. The research subjects were netizens' comments on the "Digital Era Education Development Strategy" video on Vice President @gibran_rakabuming's TikTok account. A descriptive qualitative approach was applied through the *computer-Assisted Qualitative Data Analysis* (CAQDAS) method. The entire population of 1,353 comments was collected without sampling limitations (total sampling). Data collection was conducted from the time the post was uploaded on May 2, 2025, until April 21, 2026, using a web scraping technique via Apify. The data were then analyzed using NVivo software, encompassing Sentiment Analysis and lexical extraction (Word Cloud and Word Tree). The results reveal a paradoxical reality. Quantitatively, positive sentiments dominated with an 8:3 ratio. However, the qualitative lexical analysis was predominantly filled with harsh words, personal attacks (*toxic disinhibition*), and repetitive sentence templates. This paradox proves the occurrence of public space manipulation by cyber troops (*buzzers*) to manufacture fake public consent, while organic citizen criticisms were filled with emotional aggression. This phenomenon confirms a digital civility crisis. Therefore, it is recommended that the government stop opinion manipulation practices to build authentic public trust, educational institutions reorient digital literacy curricula towards digital citizenship ethics, and platforms strengthen algorithmic transparency to crack down on *Coordinated Inauthentic Behavior*.

**Keywords:** Digital Citizenship, TikTok, Public Opinion, Toxic Disinhibition, Gibran.

***ABSTRAK***

*Penelitian ini bertujuan untuk menganalisis bentuk ekspresi dan partisipasi warga negara di ruang publik digital, serta mengungkap anomali opini publik pada unggahan kebijakan pejabat negara. Subjek penelitian adalah komentar warganet pada video "Strategi Pembangunan Pendidikan di Era Digital" di akun TikTok Wakil Presiden @gibran_rakabuming. Pendekatan kualitatif deskriptif diterapkan melalui metode Computer-Assisted Qualitative Data Analysis (CAQDAS). Data populasi sebanyak 1.353 komentar dikumpulkan secara keseluruhan (total sampling) tanpa batasan sampel. Pengumpulan data dilakukan sejak postingan tersebut diunggah pada 2 Mei 2025 hingga 21 April 2026 menggunakan teknik web scraping berbasis Apify. Data kemudian dianalisis menggunakan perangkat lunak NVivo yang mencakup Sentiment Analysis dan ekstraksi leksikal (Word Cloud dan Word Tree). Hasil penelitian*

*menunjukkan sebuah realitas paradoksal. Secara kuantitatif, sentimen positif mendominasi dengan rasio 8:3. Namun, analisis kualitatif leksikal justru didominasi oleh kata-kata kasar, serangan personal (toxic disinhibition), serta repetisi templat kalimat. Paradoks ini membuktikan terjadinya manipulasi ruang publik oleh pasukan siber (buzzer) guna memproduksi persetujuan publik palsu, sementara kritik dari warganet organik dipenuhi agresi emosional. Fenomena ini menegaskan terjadinya krisis keadaban digital (digital civility). Oleh karena itu, direkomendasikan agar pemerintah menghentikan praktik manipulasi opini demi membangun kepercayaan publik yang autentik, institusi pendidikan mereorientasi kurikulum literasi pada etika kewarganegaraan digital, dan pihak platform memperkuat transparansi algoritma untuk menindak Coordinated Inauthentic Behavior.*

***Kata kunci**: Kewarganegaraan Digital, TikTok, Opini Publik, Toxic Disinhibition, Gibran.*

## A. INTRODUCTION

The era of digital disruption has triggered a shift in TikTok's function from merely an entertainment platform to a crucial digital public space for citizens (Sukarelawati & Amalia, 2026). This platform facilitates direct and horizontal interactions between the public and public officials without the rigid bureaucratic barriers thru a more casual and humanistic communication approach (Cușnir, 2025). As a public space, TikTok provides a discursive arena that allows policies or activities of state officials to be commented on in real-time, while simultaneously breaking down traditional barriers to political participation. This phenomenon reflects the strengthening of online civic engagement, where the development of social media and the internet has redefined deliberative politics and the public sphere thru dynamic digital interactions (Mulyono et al., 2023). Nevertheless, these interactions also carry significant risks in reinforcing political polarization and identity division within society (Suffianor, 2025).

However, the openness in this digital public space poses serious challenges to communication ethics, considering that freedom of expression often exceeds the threshold of reasonableness. The use of anonymity features or pseudonymous identities on social media creates a false sense of security that significantly encourages the increase in intensity and diffusion of hate speech in the comment sections. The condition of being without face-to-face interaction triggers the effect of toxic online disinhibition, where social control weakens and personal accountability fades due to the disconnection between digital actions and real-world identities (Ritchie et al., 2026). In this dynamic, toxic online disinhibition acts as a strong moderator that connects the role of individuals as bystanders to active perpetrators of online hate (Wachs & Wright, 2018). As a result, criticism that should focus on the substance of policies often shifts to destructive digital aggression.

That bitter reality is clearly seen in the TikTok post by Vise President @gibran_rakabuming discussing "Education Development Strategies in the Digital Era." Although the content is educational and strategic for the future of the younger generation,

the comments section is flooded with irrelevant narratives. Recent studies show that exposure to such socio-political content on TikTok significantly affects students' interpersonal competence and social behavior (Robayani & Bernadus, 2026). Instead of discussing the substance of the AI technology presented, most of the comments are dominated by personal attacks and hate speech that have no correlation with digital literacy. This behavior is further exacerbated by the tendency for self-replication, where users tend to mimic toxic rhetorical patterns and negative social behaviors prevalent on the platform (Rampengan et al., 2025). This phenomenon confirms that political messages on TikTok are often distorted by audience emotions that prioritize hostile rhetoric over the substance of information (Segado-boj et al., 2026). As a result, the public education space sought thru social media has turned into an arena of digital aggression that erodes the quality of political discourse in Indonesia.

Digital citizenship competence refers to the knowledge and skills of citizens to participate actively, morally, and responsibly in the digital space, including understanding rights and responsibilities and using technology constructively (UNESCO, 2023). Digital citizenship is also understood as the effort to engage with the internet safely and with full awareness (Tadlaoui-Brahmi et al., 2022). The characteristics of this citizenship function as the main foundation of digital competence (Benaziria, 2018), which aims to foster wise behavior, communication ethics, and prevent technological crimes (Dass & Kumar, 2024). In the transition toward Society 5.0, digital citizenship becomes a mandatory pillar for character education to ensure that citizens can balance artificial intelligence with social intelligence (Yuniarto & Yudha, 2021). This competence is not limited to the technical realm but demands a deep understanding of ethics, security, and social responsibility (Mulyono et al., 2022), which are specifically divided into five pillar dimensions: digital identity, privacy management, rights and responsibilities, digital empathy, and active engagement in problem-solving (Mulyono et al., 2021).

The ethical demands are highly relevant considering the massive development of technology that requires digital citizens to possess mature literacy (Nasrulloh et al., 2017). Digital literacy is fundamentally defined as the ability to read, analyze, use, and evaluate information (Gilster, 1997), as well as the critical thinking process to evaluate received messages (Kharisma, 2017). This literacy includes the skills to access and publish information (Bawden, 2008), which enables individuals to communicate and develop creativity safely (Payton & Hague, 2010). There are four key components within it: Internet Searching for filtering accurate data (Masropah et al., 2022), Hypertextual Navigation for dynamic understanding between sources (Purba & Ain, 2024), Content Evaluation to prevent the spread of hoaxes (Amaly & Armiah, 2021), and Knowledge Assembly to integrate information (Haromain et al., 2024). Although Indonesia's digital literacy index increased to 3.65 in 2023, the comprehension rate of 62% still places Indonesia at the lowest position in the ASEAN region (Diginusa, 2024). Therefore, digital literacy education is crucial for internet users to maintain morality and national values in the online world (Putro & Tirza, 2026).

The low level of digital citizenship literacy often triggers deviant social interactions due to the Online Disinhibition Effect (ODE), where individuals feel less restrained and are bold enough to express aggression that they would not convey directly in the real world (Kreijns et al., 2004; Lapidot-Lefler & Barak, 2012). However, freedom of expression in the digital public space should still be limited by the boundaries of civic ethics (Triyanto, 2020). When this toxic disinhibition dominates, public discussion spaces are distorted into personal attacks (Meidiaputri & Mukhlis, 2023). This phenomenon has been explored by several previous studies that predominantly still use the conventional qualitative-descriptive approach (Bustami et al., 2024). For example, the research by Aser, Paramitha, and Sudarto (2022) highlights how the TikTok platform facilitates cyberbullying thru a case study interview approach. Anjani's study (2024) also uses ODE theory to explain digital aggression, but its main focus is limited to reviewing the dynamics of ITE law in Indonesia. On the other hand, Widianto's (2025) research has proven the effectiveness of sentiment analysis using automated software, but its focus remains on international geopolitical issues on the YouTube platform.

Based on the literature mapping, the research gap and novelty of this study can be emphasized. If previous studies have examined communication ethics qualitatively in a limited manner, or only conducted automatic text mining on non-public figure issues separately, then this research integrates both. This research introduces a methodological novelty by applying Computer-Assisted Qualitative Data Analysis (NVivo) software to analyze the entire population of comments without sample limitations. Specifically, this study contributes to documenting the crisis of digital civility by showing how substantive policy debates are systematically replaced by ad hominem or character assassination argument strategies (Firdaus et al., 2025). What makes this research special is its ability to empirically uncover the paradox of public opiniondemonstrating how the substance of strategic education policy is massively distorted into character assassination (ad hominem) against state officials due to the erosion of authority hierarchy in cyberspace.

The sharp contradiction between the educational message conveyed and the destructive response from the audience underscores the urgency of conducting this research. This analysis is very important considering that TikTok usage patterns have a direct and significant impact on adolescent social behavior, which can lead to social vulnerability if not guided by high ethical standards (Songgigilan et al., 2026). Low digital ethics standards can trigger threats of division in the virtual space (Relatami et al., 2026). Ultimately, this mapping of digital aggression patterns is expected not only to provide an empirical picture of netizens' political communication behavior but also to serve as a critical reflection on the quality of digital civility and the high urgency of strengthening digital citizenship competencies in Indonesia (Mulyono et al., 2021).

## B. RESEARCH METHOD

This research uses a descriptive qualitative approach that applies the Computer-Assisted Qualitative Data Analysis (CAQDAS) method. The use of CAQDAS, particularly thru

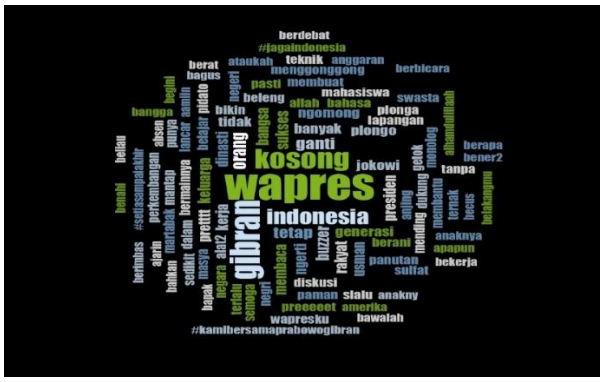
the NVivo software, enables a more structured, transparent, and integrated large-scale qualitative data analysis (Zamawe, 2015). The implementation of NVivo is crucial for managing complex data environments, providing tools for rigorous coding and mapping that ensure qualitative analysis remains systematic and verifiable (Jackson & Bazeley, 2019). The data collection process involved the extraction of the entire population, amounting to 1,353 comments from the TikTok videos under study, without any sample limitations (total sampling) to ensure a comprehensive representation of public expressions. The data collection in the form of comments was conducted from May 2, 2025, to April 21, 2026, using web scraping techniques thru the Apify application, which is highly effective for the systematic collection and preparation of social media data (Stieglitz et al., 2018). Furthermore, within this research instrument, three main indicators have been established as the priority for analysis, namely: (1) Sentiment orientation (categorized into positive, negative, and neutral sentiments); (2) Lexical patterns of digital aggression (identifying specific keywords referring to flaming and harassment); and (3) Manifestations of toxic disinhibition (focusing specifically on personal attacks or ad hominem and repetitive sentence template patterns).

The data analysis stage is fully operated using NVivo software thru two continuous approaches, starting with lexical interpretative-qualitative analysis, followed by computational analysis. In the initial stage, qualitative analysis is conducted using Word Cloud and Word Tree visualizations to map the frequency of the most dominant words that appear. After the lexical mapping is completed, the system then runs a computational Sentiment Analysis to calculate and classify the distribution of netizens' emotional polarity, visualized in the form of a Chart (sentiment graph). The sequential integration of these two NVivo features allows researchers to triangulate data to uncover anomalies in digital communication. This cross-analysis is crucial for detecting the weaknesses of the classification machine (misclassification) in reading the style of sarcasm in the initial text, as well as identifying coordinated public opinion manipulation activities (astroturfing/buzzer) behind the paradox of high positive sentiment statistics (Liu, 2012).

## C. RESULTS AND DISCUSSION

### Result

This research extracts a population of comments from video uploads related to the policy "Education Development Strategy in the Digital Era" on the TikTok account of Vise President @gibran_rakabuming. All the collected comment text data were then processed thru a preprocessing stage to clean up irrelevant symbols, emojis, and conjunctions. After being cleaned, the data was analyzed using NVivo software to visualize the frequency of the most dominant words appearing thru the Word Cloud feature. The visualization of the word frequency is as follows.

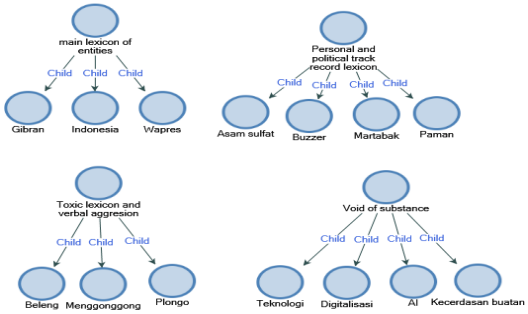



*Source: Nvivo, 2026.*

**Figure 1. Results of Word Cloud Feature Visualization**

Based on the visualization produced in Figure 1, the main lexicon dominating the comment section in terms of size and frequency consists of words referring to the subject itself and the state entity, namely "wapres," "gibran," and "indonesia." However, a crucial finding from this data processing is the massive emergence of negative, insulting, and sarcastic words. Thousands of comments are dominated by degrading words that attack the subject's intellectual capacity, such as "kosong," "plonga," "plongo," and "beleng." Furthermore, the data collection also captured high-level verbal aggression in the form of animal terminology or coarse insults, as evidenced by the prominence of the words "anjing" and "menggonggong" in the Word Cloud visualization. In addition to direct insults, the data findings also indicate a shift in sentiment toward attacks based on the subject's political track record and personal domain. This is clearly seen from the high frequency of the words "dinasti," "paman," "asam sulfat," "buzzer," and "martabak." On the contrary, the most paradoxical finding from the extraction of thousands of comments is the absence of a lexicon regarding the substance of the video. Not a single significant keyword related to "AI," "teknologi," "kecerdasan buatan," or "digitalisasi" was found to represent netizens' responses to the idea of utilizing AI presented in the video.

The data was also analyzed using the project map feature to represent the most dominant words found in the comments. The project map visualization results are as follows.

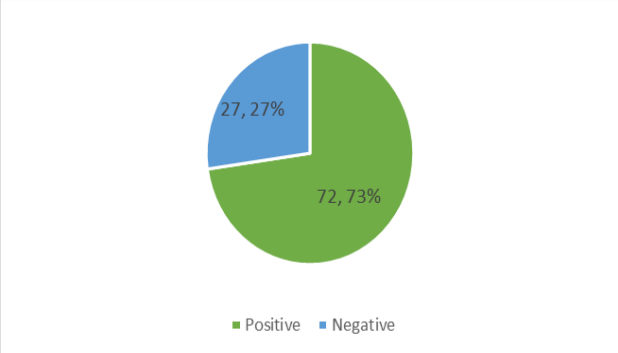



*Source: Nvivo, 2026.*

**Figure 2. Project Map Visualization Results**

Based on the Project Map visualization in Figure 2, it shows that netizens' expressions in the digital public space are more dominated by personal sentiment and

verbal aggression compared to substantial discussions about education. Although the main lexicon refers to entities such as "Gibran," "Wapres" and "Indonesia," their emergence is actually accompanied by a dense presence of toxic and degradative words like "plongo," "beleng," and "menggonggong," as well as jabs at personal political records such as "martabak," "paman," and "asam sulfat." It is very ironic when this finding also reveals a lack of lexicon relevant to the essence of the post, where there is no discourse related to "teknologi," "AI," or "digitalisasi." This indicates that the comment section fails to function as a healthy policy deliberation space, but rather is reduced to an arena of logical fallacies of the ad hominem type, where the public focuses more on attacking the character or personal track record of the figures instead of providing constructive criticism of the proposed ideas for educational development in the digital era. In addition to conducting a frequency analysis of frequently occurring words and representing those findings, the data was also analyzed using the sentiment distribution feature to compare the number of positive and negative sentiments. The sentiment classification visualization that emerged is as follows.




*Source: Processed by the researcher, 2026.*

**Figure 2. Results of Sentiment Classification Feature Visualization**

Based on the sentiment classification visualization generated thru NVivo software, the distribution of netizen responses shows a very significant proportion imbalance. Quantitatively, the graph is overwhelmingly dominated by positive sentiment at 72.73% compared to negative sentiment at 27.27%. The high frequency and volume of these positive indicators mathematically represent that the absolute majority of the total population of extracted comments are detected by the system as affirmative statements, support, or appreciation toward the subject's post. Nevertheless, the quantitative distribution figures showing the high volume of positive sentiment are purely the result of lexical data reading based on a machine dictionary by the software. The disparity between the number of positive and negative sentiments is an initial finding that necessitates researchers to conduct further qualitative and contextual data analysis to validate the authenticity of citizen interactions behind the statistical figures.

**Discussion**

The absence of lexicon related to the substance "Education Development Strategy in the Digital Era" indicates a dysfunction in the digital public space on social media in

Indonesia. The comment sections of state officials' accounts, which ideally could serve as a medium for citizen participation to discuss, engage in healthy debates, or provide constructive policy critiques, have instead been extremely distorted. That space has transformed into an arena for mass cyberbullying and a means of character assassination. Recent studies confirm that the normalization of toxic online disinhibition on social media is very evident thru the use of aggressive sarcasm, leading to the degradation of cultural values and communication ethics in the digital public sphere (Andini et al., 2025). Although the words "discussion" and "debate" appear in the Word Cloud, their very small size proves that deliberative interaction is minimal.

Netizens collectively ignore the educational value or policy information from the post. Instead of engaging in evidence-based reasoning regarding the impact of AI technology presented by the subject, social media users took shortcuts by committing the logical fallacy of ad hominem. These defensive and hostile reactions reflect a severe lack of communication ethics when facing technological discourse, which directly contradicts the ideal function of digital citizenship, which should fundamentally serve as a conceptual framework to guide ethical behavior and responsibility in navigating digital technology and artificial intelligence advancements (Prasetyo et al., 2025). They attack the messenger figure with terms like "plongo" and "kosong" and delegitimize their position by bringing up past political polarization wounds like "asam sulfat," "dinasti," and "paman" without touching on the idea being conveyed at all.

The phenomenon of using insults and animal terminology like "dog" represents the death of digital civility among some social media users. This directly reflects Indonesia's low ranking in the Digital Civility Index (DCI), where a significant portion of netizens consistently display high levels of verbal impoliteness due to the failure to internalize digital ethics within their social environment (Ramadhani & Krismono, 2023). This empirical data critically proves that the level of political participation among our netizens on short video-based platforms is still limited to purely reactionary and emotional participation. This reflects that the level of internet and social media penetration in Indonesia has not been matched by the maturity of democratic practices and adequate digital citizenship literacy.

Empirical data findings in the form of numerous comments that blatantly attack individuals confirm the applicability of the Online Disinhibition Effect theory, particularly within the spectrum of Toxic Disinhibition. The behavior recorded in the Word Cloud proves that netizens feel out of control in the online world. Harsh curse words like "beleng" and "anjing" thrown openly at a Vise President are sentences that would be almost impossible for netizens to say directly if they were face-to-face in the real world. This aggressive behavior aligns with recent findings on toxic online disinhibition in Indonesian social media, which demonstrate how psychological barriers lower online, turning digital spaces into arenas for unrestrained verbal hostility (Andini et al., 2025).

This false courage is strongly driven by the factor of Dissociative Anonymity facilitated by the TikTok platform system. This platform allows its users to create

anonymous accounts, use pseudonyms, or faceless accounts. The absence of a real identity breaks the chain of moral and social responsibility among netizens. They feel that the aggressive actions and hate speech they type will not bring legal consequences or social sanctions in their real lives, so they separate their offline identity and online behavior extremely. This psychological detachment is heavily reinforced by structural anonymity, which acts as a shield that minimizes the perceived risks of social sanctions and legal consequences for online perpetrators (Wachs & Wright, 2018).

Beside anonymity, the high frequency of insults directed at the presidential institution proves the massive occurrence of the Minimization of Authority factor. In cyberspace, netizens feel that the boundaries of hierarchy, social status, and respect for public figures become blurred and indistinct. They feel equal to the subject, which unfortunately translates to unlimited and unethical freedom of expression. Referring to the concept of digital citizenship literacy, this condition serves as an indicator that society has failed to practice the competence of Digital Empathy. When the right to comment is not accompanied by moral responsibility (Right and Responsibility), the digital space no longer becomes a pillar of democracy, but rather a medium of aggression. This decline in digital empathy directly echoes the broader crisis highlighted by Indonesia's low position in the Digital Civility Index, where online communication frequently suffers from a severe lack of respect and ethical awareness (Ramadhani & Krismono, 2023).

The numerous comments that tend to not attack the substance of the post on the TikTok account @gibran_rakabuming regarding education policy serve as evidence that the level of digital literacy among the Indonesian people is indeed still low and needs to be improved by the Indonesian society. It also requires attention from educational institutions and the government to formulate strategic policies to enhance the digital literacy of the Indonesian people. This structural demand aligns with the urgency to expand the national digital literacy program, which must systematically integrate core pillars such as digital media culture, security, and ethics to prevent further cyber misconduct (Kurniasih, 2023).

The findings of this research also reveal a paradoxical phenomenon between the results of computational sentiment extraction and lexical visualization (Word Tree). Although the Chart shows a dominance of positive sentiment with a ratio of 8:3, the Word Tree is instead dominated by a series of degradative and toxic words. This gap proves the weakness of machine classification (misclassification) against the sarcastic language style that is closely tied to Indonesian netizen culture. In natural language processing tasks, semantic ambiguity in Indonesian sarcastic utterances often causes sentiment algorithms to completely fail in recognizing true polarity, mistakenly categorizing mockery as positive sentiment (Fanani & Wahyuddin, 2026). Many comments were classified as positive sentiment by the software because they used words like "hebat" or "cerdas," which were actually forms of satire to attack the subject's capacity, such as the comment "hebatnya asam sulfat."

This anomaly reinforces the indication of cyber troop (buzzer) operations. Positive sentiment wins in quantity on the graph because it is mass-produced using short sentence

templates like "mantap," "lanjutkan." However, organic netizens who voiced criticism provided responses that were structurally longer and more descriptive, such as "dinasti," "paman," and "plongo," so these variations of negative vocabulary ultimately captured and dominated the branches in the Word Tree visualization. This shows that our digital public space is not only filled with aggression (toxic disinhibition), but has also been distorted by algorithmic manipulation (computational propaganda). This phenomenon strongly indicates the active presence of political buzzers who systematically deploy automated, coordinated inauthentic behaviors to manufacture public approval, effectively drowning out genuine organic discourse (Samad et al., 2025).

The imbalance between the number of positive and negative sentiments has both confirmed and developed the Online Disinhibition Effect theory. Theoretically, the dominance of positive sentiment with an 8:3 ratio on the graph represents the phenomenon of Benign Disinhibition. This shows that cyberspace allows citizens to express themselves more openly in providing support or hope to authority figures without the rigid constraints of hierarchy (Minimization of Authority). However, when linked to the Word Tree findings dominated by negative words, this disinhibition theory becomes distorted. Researchers argue that most of the positive sentiments are not organic disinhibition, but rather the result of Coordinated Inauthentic Behavior. In the context of digital citizenship, the factor of dissociative anonymity is no longer used solely by individuals to spontaneously release psychological barriers, but has been weaponized by cyber actors (buzzers) to create public approval by systematically deploying propaganda and manipulating digital public opinion to simulate widespread societal backing (Samad et al., 2025).

## D. CONCLUSION

Based on the analysis results using NVivo software on netizen interactions in the comment section of the TikTok account @gibran_rakabuming regarding education policies, it can be concluded that the digital public space in Indonesia currently presents a paradoxical reality and is vulnerable to manipulation. Although quantitative data shows a dominance of positive sentiment at 72.73%, in-depth lexical analysis confirms that this high figure is not a form of organic participation, but rather the result of orchestrated computational propaganda (buzzer operations) that exploit anonymity to produce manufactured consent. On the other hand, responses from organic netizens are dominated by verbal aggression, political jabs from the past, and personal attacks (toxic disinhibition) that elude detection by basic sentiment classification algorithms. This paradox clearly demonstrates the occurrence of a digital civility crisis, where the public space no longer functions as an arena for healthy and substantive democratic deliberation, but rather degrades into a battlefield between artificial narratives and citizens' emotional aggression.

Based on the data findings, the researchers provide a policy recommendation (evidence-based policy) that for political actors, it is advised to stop manipulative political communication practices thru the mobilization of cyber troops (buzzers), and return to two-way communication that prioritizes policy substance to authentically build public trust.

In the realm of education, this phenomenon signals the urgency for educational institutions to reorient digital literacy curricula, which should no longer merely focus on technical skills but be radically centered on the ethical aspects of digital citizenship, enabling society to express criticism with logical reasoning without falling into toxic behavior. Lastly, for social media platform developers (TikTok), it is crucial to enhance algorithm transparency and strengthen moderation systems to firmly address Coordinated Inauthentic Behavior, so that the organic citizen discussion ecosystem is no longer submerged by artificial voice manipulation.

## E. REFERENCES

Amaly, N., & Armiah, A. (2021). Peran kompetensi literasi digital terhadap konten hoaks dalam media sosial. *Alhadharah: Jurnal Ilmu Dakwah*, *20*(2), 43–52. https://doi.org/10.18592/alhadharah.v20i2.6019

Andini, A. V., Nurbayani, S., & Abdullah, M. N. A. (2025). Toxic online disinhibition: dampak sosial budaya penggunaan kalimat sarkasme (studi netnografi pada komunitas marah marah media sosial X). *Jurnal Ilmiah Ilmu Pendidikan*, *8*(7), 8005–8014. https://www.jiip.stkipyapisdompu.ac.id/jiip/index.php/JIIP/article/view/8520

Anjani, V. A. (2024). Cyberbullying dan dinamika hukum di Indonesia: paradoks ruang maya dalam interaksi sosial di era digital. *Jurnal Hukum Kenegaraan Dan Politik Islam*, *4*(1), 1–28. https://doi.org/10.14421/cyg94d68

Aser, F. G., Paramitha, S., & Sudarto, S. (2022). Fenomena cyberbullying di media sosial tikTok. *Kiwari*, *1*(3), 449–453. https://doi.org/10.24912/ki.v1i3.15763

Bawden, D. (2008). Origins and concepts of digital literacy. *Digital Literacies: Concepts, Policies and Practices*, *30*, 17–32.

Benaziria, B. (2018). Pengembangkan Literasi Digital pada Warga Negara Muda dalam Pembelajaran PPKn melalui Model VCT. *JUPIIS: JURNAL PENDIDIKAN ILMU-ILMU SOSIAL*, *10*, 11. https://doi.org/10.24114/jupiis.v10i1.8331

Bustami, B., Siregar, A. R., Harahap, A., & Nasution, M. S. (2024). Etika komunikasi media digital di era post-truth. *Jurnal Paradigma: Jurnal Multidisipliner Mahasiswa Pascasarjana Indonesia*, *5*(1), 39–53. https://doi.org/10.22146/jpmmpi.v5i1.91604

Cușnir, C. (2025). Public institutions meet TikTok: communication strategies and the rise of govtainment. *Media and Communication*, *13*, 1–24. https://doi.org/10.17645/mac.i496

Dass, M. A., & Kumar, M. P. (2024). Instruments for measuring Digital Citizenship Competence in schools: a scoping review. *Journal of E-Learning and Knowledge Society*, *20*(2), 9–18. https://doi.org/10.20368/1971-8829/1135934

Diginusa. (2024). *Indonesia rendah literasi digital, ini penyebabnya!* https://www.instagram.com/p/C4uvegvr68R/

Fanani, A. M., & Wahyuddin, M. I. (2026). Sarcasm detection in Indonesian YouTube comments using fine-tuned IndoBERT with class imbalance handling. *Sinkron: Jurnal Dan Penelitian Teknik Informatika Volume*, *10*(1), 26–36. https://doi.org/10.33395/sinkron.v10i1.15607

Firdaus, S., Kamal, M., & Anoegrajekti, N. (2025). Reconstruction of political ideology through ad hominem argument strategy in the 2024 Indonesian presidential-vice presidential debate. *International Seminar on Humanity, Education, and Language*, *1*(1), 35–52. https://journal.unj.ac.id/unj/index.php/ishel/article/view/57255

Gilster, P. (1997). *Digital Literacy*. Wiley. https://books.google.co.id/books?id=awkoAQAAMAAJ

Haromain, S. N., Lutfiah, L., Faxriah, A. T. N., Husen, S. A., Palupi, I. N., & Diella, D. (2024). Analisis tingkat kemampuan literasi digital siswa dalam penggunaan search engine ppplication pada pembelajaran biologi di SMAN 1 Tasikmalaya. *BIOSINTESA: Jurnal Pendidikan Biologi*, *1*(1), 1–7. https://journal.publinesia.com/index.php/biosintesa/article/view/33

Jackson, K., & Bazeley, P. (2019). *Qualitative data analysis with NVivo* (3rd ed.). SAGE Publications.

Kharisma, H. V. (2017). Literasi digital di kalangan guru SMA di Kota Surabaya. *Libri-Net*, *6*(4), 1–12.

Kreijns, K., Kirschner, P. A., Jochems, W., & van Buuren, H. (2004). Determining sociability, social space, and social presence in (a)synchronous collaborative groups. *CyberPsychology & Behavior*, *7*(2), 155–172. https://doi.org/10.1089/109493104323024429

Kurniasih, N. (2023). Digital literacy: education for safe internet usage. *Engagement: Jurnal Pengabdian Kepada Masyarakat*, *07*(1), 139–150. https://doi.org/10.29062/engagement.v7i1.1534

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Masropah, S. M., Juhanda, A., & Ramdhan, B. (2022). Analisis keterampilan literasi digital siswa SMA melalui penggunaan google lens pada konsep tumbuhan berbasis gender. *BIODIK: Jurnal Ilmiah Pendidikan Biologi*, *08*(03), 115–124. https://doi.org/10.22437/bio.v8i3.18976

Meidiaputri, R. D., & Mukhlis, I. (2023). Etika komunikasi dalam menggunakan media sosial (suatu kajian literatur ). *Cognoscere: Jurnal Komunikasi Dan Media Pendidikan*, *1*(2), 21–29. https://doi.org/10.61292/cognoscere.v1i2.71

Mulyono, B., Affandi, I., Suryadi, K., & Darmawan, C. (2021). Digital citizenship competence: initiating ethical guidelines and responsibilities for digital citizens. *ICHELSS: International Conference on Humanities, Education, Law, and Social Sciences*, *3*(1), 165–175. https://doi.org/10.62567/micjo.v3i1.1823

Mulyono, B., Affandi, I., Suryadi, K., & Darmawan, C. (2022). Online civic engagement: fostering citizen engagement through social media. *Jurnal Civics: Media Kajian Kewarganegaraan*, *19*(1), 75–85. https://doi.org/10.21831/jc.v19i1.49723

Mulyono, B., Affandi, I., Suryadi, K., & Darmawan, C. (2023). Online civic engagement through social media: an analysis of Twitter's big data. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan*, *42*(1), 12–26. https://doi.org/10.21831/cp.v42i1.54201

Nasrulloh, R., Aditya, W., Satya, T. I., Nento, M. N., Hanifah, N., Miftahussururi, M., & Akbari, Q. S. (2017). *Materi Pendukung Literasi Digital*. Kementrian Pendidikan dan Kebudayaan.

Payton, S., & Hague, C. (2010). *Digital literacy across the curriculum*. Futurelab.

Prasetyo, R. A., Mulyono, B., & Putra, W. C. (2025). Peran kewarganegaraan digital dalam pembentukan etika penggunaan artificial intelligence: systematic literature review. *Didaktika: Jurnal Kependidikan*, *14*(4), 6247–6260.

https://doi.org/10.58230/27454312.3074

Purba, A. Z., & Ain, S. Q. (2024). Peran guru dalam mengenalkan literasi digital pada siswa kelas tinggi di sekolah dasar. *Didaktika: Jurnal Kependidikan*, *13*(001), 1–10. https://doi.org/10.58230/27454312.1516

Putro, G. A., & Tirza, J. (2026). Literasi digital sebagai pilar pembentukan karakter antihoaks di era post-truth. *ARTES LIBERALES: Jurnal Ilmiah Ilmu Sosial Dan Budaya Literasi*, *2*(2), 36–44. https://ojs.uph.edu/index.php/ArtLib/article/view/10838

Ramadhani, I. F., & Krismono, K. (2023). Digital civility index in the era of information technology: a case study of higher education muslim students in Yogyakarta. *At-Thullab Jurnal: Jurnal Mahasiswa Studi Islam*, *5*(2), 253–257. https://doi.org/10.20885/tullab.vol5.iss2.art23

Rampengan, E. S., Lumapow, H. R., & Sengkey, M. M. (2025). Analisis pengaruh penggunaan media sosial tiktok terhadap perilaku imitasi diri (self imitation) pada remaja di Kota Tomohon. *Psikopedia*, *6*(1), 45–50. https://doi.org/10.53682/pj.v6i1.11575

Relatami, T., Anjelawati, D., & Kartikawati, D. (2026). Digital ethics in social media to maintain and strengthen the nation's socio-cultural values. *Multidisciplinary Indonesian Center Journal (MICJO)*, *3*(1), 768–775. https://doi.org/10.62567/micjo.v3i1.1823

Ritchie, R., Limniou, M., & Gordts, S. (2026). Greater toxic online disinhibition and lower consistent online self-presentation contribute to the perpetration of cyber dating abuse. *Frontiers in Psychiatry*, *16*, 1–13. https://doi.org/10.3389/fpsyt.2025.1716617

Robayani, F. H., & Bernadus, B. (2026). Social-political content exposure on TikTok and its impact on university students' social behavior. *International Journal of Marketing & Human Resource Research*, *7*(1), 541–557. https://doi.org/10.47747/ijmhrr.v7i1.3403

Samad, M. Y., Arsyad, A., & Kambo, G. (2025). Early detection model to prevent the damaging impact of political buzzer attacks on social media. *Otoritas: Jurnal Ilmu Pemerintahan*, *15*(3), 594–607. https://doi.org/10.26618/ojip.v15i3.17417

Segado-boj, F., Martín-quevedo, J., Gonz´alez-Aguilar, J.-M., & Antona-Jimeno, T. (2026). Political and social hate speech: differences on emotional and rhetorical features on TikTok messages. *Frontiers in Political Science*, *7*, 1–13. https://doi.org/10.3389/fpos.2025.1712927

Songgigilan, S. G., Solang, D. J., & Lovihan, M. A. K. (2026). Pengaruh penggunaan aplikasi Tiktok terhadap perilaku sosial pada remaja di Desa Sendangan Kecamatan Remboken. *Al-Zayn: Jurnal Ilmu Sosial Dan Hukum*, *4*(1), 7000–7010. https://ejournal.yayasanpendidikandzurriyatulquran.id/index.php/AlZayn/article/view/4480

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Suffianor, S. (2025). Social media, populism, and identity in shaping political polarization. *Politeia: Journal of Public Administration and Political Science and International Relations*, *3*(2), 72–84. https://doi.org/10.61978/politeia.v3i2.966

Sukarelawati, S., & Amalia, N. (2026). Analysis of @xproject.id's TikTok political communication strategies via the SOSTAC model. *Jurnal Studi Komunikasi*, *10*(1), 305–318. https://doi.org/10.25139/jsk.v10i1.11106

Tadlaoui-Brahmi, A., Çuko, K., & Alvarez, L. (2022). Digital citizenship in primary education:

A systematic literature review describing how it is implemented. *Social Sciences and Humanities Open*, *6*(1), 1–9. https://doi.org/10.1016/j.ssaho.2022.100348

Triyanto, T. (2020). Peluang dan tantangan pendidikan karakter di era digital. *Jurnal Civics: Media Kajian Kewarganegaraan*, *17*(2), 175–184. https://doi.org/10.21831/jc.v17i2.35476

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific, and Cultural Organization.

Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: the moderating role of toxic online disinhibition. *International Journal of Environment Research and Public Health*, *15*(2030), 1–9. https://doi.org/10.3390/ijerph15092030

Widianto, F. (2025). Analisis sentimen komentar Youtube tentang konflik Iran- Israel menggunakan orange data mining. *Sains Data Jurnal Studi Matematika Dan Teknologi*, *3*(2), 81–88. https://doi.org/10.52620/sainsdata.v3i2.278

Yuniarto, B., & Yudha, R. P. (2021). Literasi digital sebagai penguatan pendidikan karakter menuju era society 5.0. *Edueksos: Jurnal Pendidikan Sosial Dan Ekonomi*, *10*(2), 176–194. https://doi.org/10.24235/edueksos.v10i2.8096

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: evidence-based reflections. *Malawi Medical Journal*, *27*(1), 13–15. https://doi.org/10.4314/mmj.v27i1.4