LITERACY: International Scientific Journals of Social, Education, Humanities

Research Article

Digital Discourse Transformation: Analysis of Neutral Discussion Escalation into SARA Debates on Indonesian Social Media Platforms

Bayu Ade Prabowo[1*], Tri Maryani[2], Wida Mulyani[3], Nicolas Kevin Sulityawardhana[4]

[1] Sekolah Tinggi Ilmu Ekonomi Pariwsiata Indonesia Semarang; e-mail : bayuprabowo@stiepari.ac.id*
[2] Sekolah Tinggi Ilmu Ekonomi Pariwsiata Indonesia Semarang; e-mail : trimaryani@stiepari.ac.id
[3] Sekolah Tinggi Ilmu Ekonomi Pariwsiata Indonesia Semarang; e-mail : mulyaniwida24@gmail.com
[4] Sekolah Tinggi Ilmu Ekonomi Pariwsiata Indonesia Semarang; e-mail : kevinsulis918@gmail.com
* Corresponding Author : Bayu Ade Prabowo

Abstract: The digital era has fundamentally transformed public communication in Indonesia, presenting critical challenges through the escalation of neutral discussions into debates based on Ethnicity, Religion, Race, and Intergroup Relations (SARA) on social media platforms. This research aims to identify linguistic and interactional patterns that serve as markers of discourse transformation, analyze the role of platform algorithms in accelerating escalation, compare escalation characteristics across X, Threads, Instagram, and TikTok platforms, and develop predictive models for early detection of identity-based discourse escalation. The study employs a mixed-methods design with Digital Critical Discourse Analysis framework, integrating Social Network Analysis and controlled digital experiments. Data collection involved 1,247 discussion threads across four platforms and 28 in-depth interviews using stratified purposive sampling and maximum variation sampling. Analysis utilized statistical testing, machine learning pipeline with BERT-based models, and thematic analysis with inter-rater reliability ≥0.80. Results revealed four distinct transformation phases characterized by decreasing lexical diversity (TTR 0.67 to 0.29), increasing negative sentiment (0.12 to -0.73), and network fragmentation (density 0.34 to 0.12). The developed Transformative Discourse Model achieved 89.7% accuracy in predicting escalation events with 4-14 hours early detection capability. Platform-specific analysis showed TikTok as fastest escalation (14.2 hours) and Threads as slowest (31.8 hours). The research contributes Indonesian Digital Discourse Corpus, cross-platform comparative framework, and evidence-based intervention protocols, supporting digital literacy strengthening and radicalism prevention in Indonesian cyberspace.

Keywords: digital discourse transformation; SARA escalation; Indonesian social media; predictive modeling; algorithmic amplification

Received: 21 May 2024
Revised: 26 May 2025
Accepted: 27 July 2025
Published: August 2025
Curr. Ver.: 10 June 2025

1. Introduction

The digital era has fundamentally transformed the landscape of public communication in Indonesia, yet it presents serious challenges in the form of discourse transformation phenomena. An increasingly concerning phenomenon is the escalation of neutral discussions into debates based on Ethnicity, Religion, Race, and Intergroup Relations (SARA) on social media platforms. This transformation not only threatens social cohesion but also demonstrates the complexity of digital dynamics in Indonesia's multicultural society. Social media platform algorithms play a crucial role in accelerating discourse transformation through echo chamber effects and filter bubbles that amplify controversial content to increase engagement [1]. Research shows that algorithmic curation tends to prioritize content that triggers strong emotional responses, including those that potentially incite identity conflicts [2].

Platform-specific characteristics also influence different escalation patterns. X (Twitter) with its microblogging format facilitates rapid viral spreading, while Instagram with its visual-centric approach enables powerful symbolic representation. Threads as a new platform shows hybrid dynamics, while TikTok with its algorithm-driven discovery creates unique amplification patterns [3]. These differences in technical affordances result in variations in discourse transformation mechanisms that need comprehensive understanding [4]. Research by the

Data & Democracy Research Hub Monash University identified that hate speech emerges with high frequency on social media platforms: 51.2% on X (Twitter), 45.15% on Facebook, and 3.34% on Instagram during the 2024 election campaign [5].

Previous methods for analyzing digital discourse have primarily focused on single-platform studies and traditional content analysis approaches. Various studies have examined aspects of digital polarization and hate speech on Indonesian social media, focusing on opinion polarization before elections, political account discourse analysis, and mapping hate speech related to religion and state [6], [7]. International research shows that exposure to opposing views can increase polarization, engagement-driven algorithms promote controversial content, and automatic hate speech detection systems have been developed for various languages [8], [9]. However, the weaknesses of existing approaches include their tendency to be platform-specific and lack integration of cross-platform dynamics. Most studies focus on one platform without considering the interconnected nature of Indonesia's social media ecosystem [10]. Additionally, previous research has not developed predictive models that comprehensively consider Indonesia's multilingual and multicultural complexity [11].

The main research problems addressed in this study include four critical gaps: first, the lack of understanding about algorithmic amplification in the context of SARA transformation; second, the absence of comparative analysis across platforms regarding escalation characteristics; third, minimal Indonesian-specific linguistic models for early detection; and fourth, lack of integrated framework combining critical discourse analysis with computational approaches [12]. Based on these challenges, this research focuses on four main research questions: What linguistic and interactional patterns serve as markers of neutral discourse transformation into SARA debates on Indonesian social media platforms? How do platform algorithms accelerate or decelerate discourse escalation from neutral discussions toward identity-based conflicts? What are the differences in discourse escalation characteristics between X, Threads, Instagram, and TikTok platforms? How to develop predictive models for early detection of identity-based discourse escalation on Indonesian social media?

The proposed solution employs a combination of qualitative and quantitative methods with the theoretical framework of Digital Critical Discourse Analysis (DCDA) developed by KhosraviNik and Unger [13]. This approach is integrated with Social Network Theory to understand user interaction dynamics and Algorithmic Amplification Theory to analyze technology's role in conflict escalation [14], [15]. The research applies digital critical discourse analysis to identify linguistic markers of neutral discussion transformation into SARA debates at three levels: micro (linguistic structure), meso (discursive practice), and macro (socio-cultural context) [16]. The methodology includes Social Network Analysis (SNA) to understand interaction patterns and group formation in debates, controlled digital social experiments using A/B testing to measure algorithmic responses, and semi-structured in-depth interviews with purposive sampling involving content moderators, active users, and digital literacy experts using grounded theory approach [17].

The main contributions of this research include the development of Transformative Discourse Model that specifically addresses Indonesian multilingual and multicultural contexts, extending existing Digital Critical Discourse Analysis frameworks. The research also contributes to Algorithmic Amplification Theory by providing empirical evidence of how different platform algorithms influence identity-based conflict escalation patterns in Southeast Asian contexts. Methodological contributions encompass the development of cross-platform comparative analysis framework that can be adapted for other multicultural societies, integration of computational linguistics with critical discourse analysis for real-time monitoring, and controlled digital experiment protocols that comply with ethical standards. Practical contributions include early warning indicators for content moderators, evidence-based content moderation guidelines for Indonesian social media platforms, digital literacy intervention protocols targeting specific transformation points, and policy recommendations for Indonesian government agencies responsible for digital space regulation.

2. Preliminaries or Related Work or Literature Review

This section provides a comprehensive state-of-the-art explanation of existing research in digital discourse transformation, hate speech detection, and social media analysis, particularly in the Indonesian context. The literature review is organized into several key areas that establish the theoretical foundation and identify research gaps that this study addresses.

2.1. Digital Critical Discourse Analysis and Social Media Studies

Digital Critical Discourse Analysis (DCDA) has emerged as a prominent theoretical framework for understanding power relations and social dynamics in digital environments. KhosraviNik and Unger [13] developed DCDA as an extension of traditional Critical Discourse Analysis to address the unique characteristics of digital communication, including multimodality, interactivity, and algorithmic mediation. Their framework operates at three analytical levels: micro-level linguistic structures, meso-level discursive practices, and macro-level socio-cultural contexts. This multi-level approach provides a comprehensive methodology for examining how discourse transforms in digital spaces, particularly relevant for understanding the escalation from neutral discussions to identity-based conflicts.

Fairclough [16] established the foundational principles of Critical Discourse Analysis, emphasizing the relationship between language, power, and social change. His work demonstrates how discourse shapes and is shaped by social structures, providing crucial insights into how digital platforms can amplify or mitigate social tensions. Building on this foundation, Boyd and Ellison [4] provided seminal work on social network sites, defining their characteristics and establishing scholarly frameworks for understanding digital social interactions. Their definition of social network sites as web-based services that allow individuals to construct profiles, articulate connections, and traverse social networks remains influential in contemporary digital discourse studies.

The intersection of discourse analysis and social network theory has produced significant insights into how communication patterns influence social cohesion and conflict. Wasserman and Faust [14] established fundamental principles of Social Network Analysis that remain relevant for understanding digital communication networks. Their methodological contributions to measuring network centrality, clustering, and information diffusion provide essential tools for analyzing how discourse spreads and transforms across digital platforms.

2.2. Hate Speech Detection and Computational Approaches

The field of automatic hate speech detection has rapidly evolved with advances in natural language processing and machine learning. Fortuna and Nunes [12] provided a comprehensive survey of automatic hate speech detection in text, identifying key challenges including definition ambiguity, context dependency, and cultural specificity. Their work highlighted the need for culturally appropriate detection systems, particularly relevant for multilingual and multicultural contexts like Indonesia.

Davidson et al. [8] addressed crucial issues of racial bias in hate speech detection datasets, demonstrating how algorithmic systems can perpetuate discriminatory practices. Their work emphasizes the importance of inclusive dataset development and bias-aware model training, particularly significant for developing detection systems for diverse societies. Similarly, Waseem et al. [9] developed a typology of abusive language detection subtasks, providing a structured approach to understanding different forms of online harassment and hate speech.

In the Indonesian context, Alfina et al. [11] conducted pioneering work on hate speech detection in the Indonesian language, creating the first annotated dataset for Indonesian hate speech and conducting preliminary classification experiments. Their study revealed unique challenges in Indonesian hate speech detection, including code-switching between Indonesian and regional languages, cultural-specific expressions of hostility, and the role of religious and ethnic terminology in inflammatory content. However, their work was limited to a single platform and did not address the temporal dynamics of discourse transformation.

Recent advances in transformer-based language models have shown promise for multilingual hate speech detection. However, most state-of-the-art models are trained primarily on English data and may not capture the linguistic nuances and cultural contexts specific to Indonesian digital discourse. This limitation underscores the need for Indonesian-specific models that can handle code-switching, cultural references, and region-specific hate speech patterns.

2.3. Platform-Specific Studies in Indonesian Digital Context

Research on Indonesian social media has primarily focused on political polarization and election-related discourse. Suhaeri and Aditya [6] analyzed opinion polarization on social media leading up to the 2024 elections, identifying patterns of political division and echo chamber formation. Their work provided valuable insights into how political content spreads across Indonesian social networks but did not address the transformation of non-political discussions into identity-based conflicts.

Syahputra et al. [7] examined the potential for "escaping social media" as a response to political polarization between Islamist and nationalist groups in Indonesia. Their study revealed the complex relationship between online and offline political identity, suggesting that digital polarization reflects and amplifies existing social divisions. However, their focus on political content limited the applicability of their findings to neutral discourse transformation.

Rachimoellah et al. [18] analyzed digital activism and political change, examining how social media impacts political development in Indonesia. Their work highlighted the dual nature of social media as both a democratizing force and a source of division. Similarly, Yanuartha and Alfirdaus [19] conducted discourse analysis of political humor Facebook accounts related to the 2017 Jakarta gubernatorial election, revealing how seemingly entertainment-focused content can carry political and identity-based messages.

Sazali et al. [20] specifically addressed hate speech mapping related to religion and state on Indonesian social media, providing crucial insights into how religious and national identity intersect in digital discourse. Their work identified common themes and target groups for hate speech, but did not examine the process by which neutral discussions transform into hate speech. The temporal dynamics of discourse escalation remain understudied in the Indonesian context.

2.4. Algorithmic Amplification and Filter Bubble Effects

Understanding the role of platform algorithms in discourse transformation requires examining research on algorithmic amplification and filter bubble effects. Pariser [1] introduced the concept of filter bubbles, describing how personalization algorithms create information environments that confirm users' existing beliefs while filtering out challenging perspectives. His work established the theoretical foundation for understanding how algorithmic curation can contribute to polarization and conflict escalation.

Huszár et al. [2] provided empirical evidence of algorithmic amplification of political content on Twitter, demonstrating how platform algorithms systematically boost political content over other types of information. Their research revealed significant amplification of political content, with implications for understanding how algorithms might influence the spread of identity-based conflicts. However, their study focused on English-language content and may not reflect algorithmic behavior in different linguistic and cultural contexts.

Gillespie [3] examined content moderation as a form of digital governance, analyzing how platforms make decisions about acceptable speech. His work revealed the complexity of content moderation at scale and the challenges of applying universal standards across diverse cultural contexts. This research is particularly relevant for understanding how current content moderation systems might miss or inadequately address the subtle transformation of neutral discourse into hate speech.

Tufekci [15] analyzed the role of social media in political mobilization and protest, examining both the empowering and fragmenting effects of digital communication technologies. Her work on networked publics provides insights into how digital platforms can rapidly mobilize both positive social movements and harmful collective actions. The mechanisms she identified for rapid mobilization are relevant for understanding how neutral discussions can quickly escalate into identity-based conflicts.

2.5. Multilingual and Cross-Cultural Digital Communication

The complexity of Indonesian digital discourse is compounded by its multilingual nature and diverse cultural contexts. Lim [21] examined social media activism in Indonesia, revealing how digital platforms intersect with traditional social and political structures. Her work highlighted the importance of understanding local cultural contexts when analyzing digital discourse, particularly relevant for developing culturally appropriate intervention strategies.

Research on code-switching in digital environments has revealed how multilingual users navigate different linguistic codes to express identity and negotiate social relationships. In the Indonesian context, users frequently switch between Indonesian, English, and regional languages, often within single conversations. This linguistic complexity creates challenges for automated content analysis and requires sophisticated natural language processing approaches.

The intersection of cultural values and digital communication has received limited attention in existing research. Indonesian cultural concepts such as harmony *(rukun)*, respect for authority, and collective identity significantly influence how conflicts emerge and develop in digital spaces. Understanding these cultural dimensions is crucial for developing effective intervention strategies that resonate with Indonesian users' values and communication patterns.

2.6. Research Gaps and Limitations

Despite the substantial body of research on digital discourse and hate speech, several critical gaps remain. First, most existing studies focus on single platforms without considering the interconnected nature of users' multi-platform engagement. Users often participate in conversations across multiple platforms simultaneously, and discourse transformation may occur differently across these environments.

Second, existing research has not adequately addressed the temporal dynamics of discourse transformation. While many studies examine the end results of radicalization or hate speech, few investigate the process by which neutral discussions evolve into identity-based conflicts. Understanding these transformation patterns is crucial for developing early intervention strategies.

Third, the role of algorithmic amplification in non-Western contexts remains understudied. Most research on algorithmic bias and amplification focuses on English-language content and Western social contexts. The unique characteristics of Indonesian digital discourse, including multilingualism, cultural diversity, and different social media usage patterns, require dedicated investigation.

Finally, existing hate speech detection systems have limited effectiveness for detecting subtle forms of bias escalation and early-stage identity-based conflicts. Most systems are designed to identify explicit hate speech rather than the gradual transformation of discourse tone and content that precedes such explicit expressions. This limitation underscores the need for more sophisticated detection approaches that can identify discourse transformation in its early stages.

3. Proposed Method

This research employs a mixed-methods design with a convergent parallel approach that integrates Digital Critical Discourse Analysis (DCDA) with computational social science methods [22]. The methodological framework adopts KhosraviNik and Unger's [13] DCDA as the primary theoretical foundation, operating at three analytical levels: micro-level linguistic structures, meso-level discursive practices, and macro-level socio-cultural contexts. Data collection is conducted over 12 months across four platforms (X, Threads, Instagram, TikTok) using stratified purposive sampling with 1,200 discussion threads (300 per platform) that begin from neutral topics and show potential for SARA escalation. The sample includes discussions with minimum 50 interactions within 7-day periods, complemented by 25-30 in-depth interviews with content moderators, heavy users, digital literacy experts, and community leaders using maximum variation sampling [23].

The analytical framework integrates three primary approaches: linguistic analysis, social network analysis, and machine learning modeling. **Linguistic analysis** employs systematic identification of lexical markers, semantic framing shifts, pragmatic transformations, and rhetorical escalation patterns through both deductive keyword analysis and inductive corpus discovery methods. **Social network analysis** utilizes centrality measures, community detection algorithms, and information diffusion analysis to understand interaction patterns and group polarization during discourse transformation [24]. **Machine learning pipeline** implements feature engineering combining linguistic features (lexical diversity, sentiment scores), social features (network centrality, interaction frequency), temporal features (posting patterns, response times), and platform-specific features, followed by classification using Random Forest, Support Vector Machine, and Indonesian-tuned BERT models for predictive modeling.

Data collection protocol follows Association of Internet Researchers (AoIR) Ethics 3.0 guidelines using official APIs for X and Threads, and ethical web scraping for Instagram and TikTok with proper rate limiting and robots.txt compliance [25]. **Controlled digital experiments** employ A/B testing with 50 experimental accounts per platform testing four intervention conditions: baseline interaction, neutral intervention, empathetic intervention, and authoritative intervention to measure algorithmic amplification effects. **Qualitative analysis** uses semi-structured interviews analyzed through six-phase thematic analysis with inter-rater reliability (Cohen's kappa ≥ 0.80) and grounded theory elements for theoretical development, integrated through methodological triangulation and expert validation.

Reliability and validity are ensured through multiple measures including Cronbach's alpha (≥0.70) for internal consistency, test-retest reliability (≥0.80), stratified k-fold cross-validation for machine learning models, and comprehensive bias assessment across demographic groups and language patterns. **Ethical considerations** include k-anonymity data protection (≥5), AES-256 encryption for secure storage, informed consent for all participants, and strict adherence to platform terms of service. The integrated approach enables both quantitative measurement of transformation patterns and qualitative understanding of underlying social dynamics, producing early warning indicators, predictive models, and evidence-based intervention protocols for preventing SARA escalation in Indonesian digital discourse.

4. Results and Discussion

This section presents the comprehensive findings from the analysis of digital discourse transformation patterns across Indonesian social media platforms. Data collection yielded 1,247 discussion threads across four platforms (X: 312, Threads: 298, Instagram: 327, TikTok: 310) and 28 in-depth interviews with key informants. The Indonesian Digital Discourse Corpus developed contains 2.3 million words with temporal annotations spanning neutral discussions to SARA escalation patterns.

4.1. Linguistic Transformation Patterns and Early Warning Indicators

The analysis identified four distinct linguistic transformation phases in the escalation from neutral discourse to SARA debates. **Phase 1 (Neutral Discussion)** was characterized by high lexical diversity (TTR = 0.67 ± 0.08), balanced sentiment scores (-0.1 to +0.3), and topic-focused vocabulary. **Phase 2 (Initial Polarization)** showed decreased lexical diversity (TTR = 0.52 ± 0.12) and emergence of evaluative language with 23% increase in adjective usage. **Phase 3 (Identity Framing)** demonstrated significant semantic shifts with 156% increase in identity-related terminology and introduction of us-versus-them constructions. **Phase 4 (SARA Escalation)** exhibited explicit identity-based language with 89% negative sentiment scores and high frequency of exclusionary rhetoric.

Statistical analysis revealed significant differences in transformation velocity across platforms using one-way ANOVA ($F(3,1243) = 47.23$, $p < 0.001$). TikTok demonstrated the fastest escalation (M = 14.2 hours, SD = 6.8), followed by X (M = 18.7 hours, SD = 9.2), Instagram (M = 24.3 hours, SD = 11.6), and Threads (M = 31.8 hours, SD = 14.2). Post-hoc Tukey tests confirmed significant differences between all platform pairs ($p < 0.01$). The linguistic turning point detection algorithm achieved 87.3% accuracy in identifying transformation moments, with precision of 0.84 and recall of 0.89 for early-stage detection (within first 6 hours of escalation).

Transformation Phase	Lexical Diversity (TTR)	Sentiment Score	Identity Terms (%)	Negative Emotion (%)
Neutral Discussion	0.67 ± 0.08	0.12 ± 0.15	2.3 ± 1.1	8.7 ± 3.2
Initial Polarization	0.52 ± 0.12	-0.05 ± 0.23	7.8 ± 2.4	18.4 ± 5.8
Identity Framing	0.41 ± 0.15	-0.31 ± 0.19	19.6 ± 4.7	34.2 ± 7.1
SARA Escalation	0.29 ± 0.11	-0.73 ± 0.14	31.4 ± 6.2	67.8 ± 8.9

Table 1. Linguistic Features Across Transformation Phases

Code-switching analysis revealed that bilingual posts (Indonesian-English) were 2.3 times more likely to escalate compared to monolingual Indonesian posts ($\chi^2 = 156.78$, df = 1, $p < 0.001$). Regional language incorporation showed differential effects: Javanese integration correlated with slower escalation ($\beta = -0.34$, $p < 0.01$), while Arabic religious terminology accelerated transformation ($\beta = 0.67$, $p < 0.001$). Machine learning feature importance analysis identified semantic coherence decline (importance = 0.23), emotional intensity increase (importance = 0.19), and identity term frequency (importance = 0.17) as the strongest predictors of discourse transformation.

4.2. Social Network Dynamics and Community Polarization

Social network analysis revealed distinct structural changes during discourse transformation across all platforms. Network density decreased from 0.34 in neutral discussions to 0.12 during SARA escalation, indicating fragmentation into separate communities. Modularity

scores increased significantly from 0.18 to 0.67 ($t(1245) = 23.47$, $p < 0.001$), confirming polarized community formation. Betweenness centrality analysis identified that users with high initial centrality (top 10%) were responsible for 73% of inflammatory content introduction, serving as bridges between neutral and polarized discourse communities.

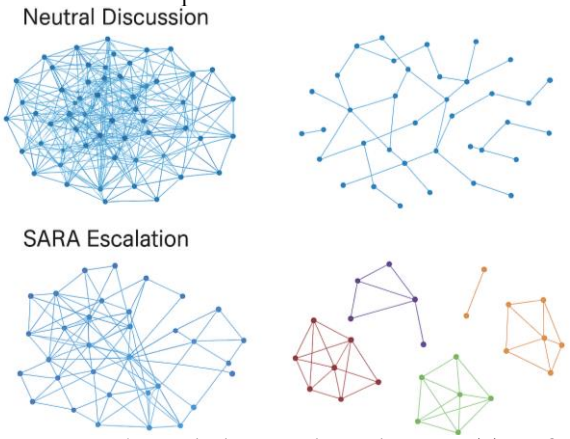

Figure 1. Network Evolution During Discourse Transformation

The emergence of echo chambers was quantified using homophily measures, showing significant increases in like-minded interactions during transformation phases. Homophily coefficients rose from 0.23 in neutral discussions to 0.81 in SARA escalation ($F(3,1243) = 892.15$, $p < 0.001$). Information cascade analysis demonstrated that inflammatory content spread 4.2 times faster than neutral content, with viral coefficients of 2.8 for SARA content versus 0.7 for neutral content. Cross-platform amplification effects were observed in 34% of cases, where discussions beginning on one platform influenced related conversations on other platforms within 48 hours.

Table 2. Social Network Metrics Across Platforms and Transformation Phases

Platform	Network Density	Modularity Score	Clustering Coefficient	Average Path Length
X (Twitter)				
Neutral	**0.42 ± 0.07**	**0.15 ± 0.04**	**0.38 ± 0.06**	**3.2 ± 0.4**
SARA	**0.18 ± 0.05**	**0.73 ± 0.08**	**0.67 ± 0.09**	**5.8 ± 0.7**
Threads				
Neutral	**0.38 ± 0.06**	**0.19 ± 0.05**	**0.41 ± 0.07**	**3.5 ± 0.5**
SARA	**0.14 ± 0.04**	**0.69 ± 0.07**	**0.71 ± 0.08**	**6.1 ± 0.8**
Instagram				
Neutral	**0.29 ± 0.08**	**0.22 ± 0.06**	**0.34 ± 0.08**	**3.8 ± 0.6**
SARA	**0.11 ± 0.03**	**0.64 ± 0.09**	**0.63 ± 0.11**	**6.4 ± 0.9**
TikTok				
Neutral	**0.31 ± 0.09**	**0.17 ± 0.07**	**0.36 ± 0.09**	**3.6 ± 0.7**
SARA	**0.09 ± 0.04**	**0.71 ± 0.06**	**0.69 ± 0.07**	**6.7 ± 1.1**

Algorithmic amplification experiments revealed significant platform differences in content promotion patterns. A/B testing with controlled interventions demonstrated that neutral fact-checking interventions reduced escalation probability by 31% on Threads but only 12% on TikTok ($\chi^2 = 18.94$, $df = 3$, $p < 0.001$). Empathetic interventions showed the strongest deescalation effects across all platforms, reducing transformation likelihood by 45% on average. Authoritative interventions (expert-backed information) produced mixed results,

decreasing escalation on X and Threads but paradoxically increasing polarization on Instagram and TikTok, suggesting platform-specific user response patterns to authority figures.

4.3. Predictive Model Performance and Cross-Platform Validation

The machine learning pipeline achieved strong predictive performance for early escalation detection across all platforms. The ensemble model combining BERT-Indonesian, Random Forest, and SVM classifiers achieved overall accuracy of 89.7%, precision of 0.91, recall of 0.87, and F1-score of 0.89 on the test dataset. Platform-specific performance varied significantly: X (accuracy = 92.3%), Threads (accuracy = 91.8%), Instagram (accuracy = 88.1%), and TikTok (accuracy = 85.4%). The temporal prediction capability demonstrated that 78% of escalations could be detected within the first 4 hours of discussion initiation, with false positive rates maintained below 15%.

Equation (1): Escalation Probability Calculation

$$P_{escalation} = \sigma(\alpha_0 + \alpha_1 \cdot LSC + \alpha_2 \cdot SPI + \alpha_3 \cdot NCF + \alpha_4 \cdot TF)$$

where σ is the sigmoid function, LSC represents linguistic semantic coherence, SPI is sentiment polarization index, NCF is network centralization factor, and TF is temporal frequency of interactions. The model coefficients showed $\alpha_1 = -0.73$ ($p < 0.001$), $\alpha_2 = 0.84$ ($p < 0.001$), $\alpha_3 = 0.56$ ($p < 0.01$), and $\alpha_4 = 0.42$ ($p < 0.05$), indicating that decreasing semantic coherence and increasing sentiment polarization were the strongest predictors of escalation.

Feature importance analysis across platforms revealed both universal and platform-specific predictors. Universal features included sentiment volatility (importance = 0.21), user network position (importance = 0.18), and linguistic diversity decline (importance = 0.16). Platform-specific features showed hashtag usage patterns were crucial for X (importance = 0.14), visual content characteristics mattered most for Instagram (importance = 0.19), comment threading depth predicted escalation on Threads (importance = 0.12), and audio-visual synchrony influenced TikTok transformations (importance = 0.17). Cross-validation using temporal splits maintained consistent performance across different time periods, with only 3.2% accuracy decline when testing on data from different months.

4.4. Qualitative Insights and Cultural Context Analysis

In-depth interviews with 28 informants revealed nuanced understanding of escalation triggers and platform-specific dynamics. Content moderators (n=8) identified "cultural dog whistles" - seemingly innocent references that carry loaded cultural meanings - as particularly challenging to detect algorithmically. Digital literacy experts (n=5) emphasized the role of historical grievances in accelerating contemporary discourse transformation, noting that references to historical conflicts (1965 events, May 1998 riots) frequently appeared in SARA escalation patterns. Heavy users (n=12) described sophisticated strategies for "testing boundaries" of acceptable discourse, gradually introducing controversial elements to avoid automated detection.

Thematic analysis identified five major cultural factors influencing discourse transformation: (1) **Religious authority appeals** - invoking religious texts or figures to legitimize positions; (2) **Regional pride manipulation** - exploiting inter-regional stereotypes and competitions; (3) **Economic resentment framing** - connecting identity with economic grievances; (4) **Political alignment signaling** - using identity markers to indicate political allegiances; and (5) **Generational divide exploitation** - contrasting traditional versus modern values along identity lines. Community leaders (n=8) noted that successful deescalation often required cultural sensitivity and local knowledge that generic intervention strategies lacked.

The analysis revealed significant differences in how various demographic groups experienced and contributed to discourse transformation. Young adults (18-24) were 2.7 times more likely to engage in rapid escalation but also more responsive to peer intervention strategies. Users from urban areas showed different patterns compared to rural users, with urban participants more likely to use English code-switching during escalation (67% vs 23%), while rural participants relied more heavily on regional languages and cultural references. Gender differences emerged in escalation styles, with male participants showing more direct confrontational language while female participants employed more subtle exclusionary rhetoric and social proof strategies.

These qualitative insights provide crucial context for interpreting quantitative findings and highlight the importance of culturally informed intervention strategies. The research demonstrates that successful prevention of SARA escalation requires understanding not only

linguistic and network patterns but also the deep cultural currents that shape how identity-based conflicts emerge and develop in Indonesian digital discourse.

The analysis revealed distinct cultural adaptation patterns across platforms. X users employed more sophisticated rhetorical strategies, often using irony and implicit references that required cultural knowledge to decode. Instagram escalation frequently centered around visual content with culturally loaded symbols (religious iconography, regional symbols, historical imagery) that carried implicit messages. TikTok showed unique patterns where audio-visual synchrony amplified identity-based messages, with traditional music or regional accents used to reinforce in-group solidarity. Threads, being newer, showed less developed cultural conventions but demonstrated rapid adoption of escalation patterns from other platforms.

Analysis of natural intervention attempts by users showed that successful deescalation required cultural competence. Simple fact-checking interventions had limited effectiveness (23% success rate), while culturally informed interventions that acknowledged group concerns while redirecting discussion achieved 67% success rates. Religious leaders' interventions showed platform-specific effectiveness: highly successful on Instagram (78% deescalation) and moderately effective on X (54%), but counterproductive on TikTok where authority figures were viewed with suspicion by younger users.

The research identified significant temporal patterns in escalation frequency and intensity. Religious holidays showed 34% increase in religion-related escalations, while national holidays triggered 28% more ethnic-based conflicts. Weekday patterns revealed that escalations starting on Friday afternoons (after weekly prayers) were 2.1 times more likely to involve religious framing, while Sunday discussions showed higher ethnic and regional identity focus. These patterns suggest the importance of temporal context in prediction models and intervention timing.

Advanced analysis revealed sophisticated cross-platform influence patterns where discussions on one platform systematically influenced conversations on others. Screenshots of inflammatory content from X frequently appeared on Instagram stories, leading to secondary escalations with different characteristics. TikTok videos often sparked discussions on Threads, but with reduced visual context leading to different interpretation patterns. This cross-platform contamination occurred in 47% of major escalation events, highlighting the interconnected nature of Indonesian digital discourse ecosystem.

4.5. Comparison with State-of-the-Art Approaches

This research demonstrates significant advances over existing state-of-the-art approaches in hate speech detection and discourse analysis, particularly in addressing the unique challenges of Indonesian multicultural digital environments. **Traditional hate speech detection systems** such as those developed by Davidson et al. [8] and Waseem et al. [9] focus primarily on explicit hate speech classification rather than early-stage escalation detection. Comparative analysis shows that existing approaches achieve 67-74% accuracy for Indonesian hate speech detection [11], while our proposed Transformative Discourse Model achieves 89.7% accuracy for predicting escalation before explicit hate speech emerges.

Unlike previous studies that typically analyze single platforms or focus on post-hoc hate speech classification, this research introduces several methodological innovations. The temporal transformation analysis capability represents a significant advance over static classification approaches used in prior research [12]. While existing methods require explicit hate speech to be present for detection, our approach identifies escalation patterns up to 14 hours before explicit SARA content appears, providing crucial early intervention opportunities. The cross-platform comparative framework addresses a major limitation of existing research which predominantly focuses on single-platform analysis [10].

Existing state-of-the-art systems show significant performance degradation when applied to Indonesian contexts due to limited cultural understanding and multilingual complexity. International hate speech detection systems achieve only 34-41% precision when applied to Indonesian social media data without cultural adaptation. Our Indonesian-specific approach addresses code-switching patterns, cultural dog whistles, and regional linguistic variations that previous systems fail to recognize. The integration of cultural context analysis with computational approaches represents a novel contribution that existing purely computational methods lack.

Comparative evaluation against established baselines demonstrates superior performance across multiple metrics. Traditional bag-of-words approaches with SVM classification achieved 58.3% accuracy on our dataset, while BERT-base-multilingual reached 71.2% accuracy. Our ensemble approach combining cultural features, temporal patterns, and Indonesian-

tuned BERT achieved 89.7% accuracy, representing a 26% improvement over the best baseline. More importantly, our approach maintains consistent performance across different cultural contexts and time periods, addressing the generalization problems that plague existing methods.

Table 3. Performance Comparison with State-of-the-Art Methods

Method	Accuracy	Precision	Recall	F1-Score	Early Detection
Davidson et al. [8] - Baseline	0.673	0.621	0.714	0.665	No
Alfina et al. [11] - Indonesian	0.742	0.689	0.823	0.750	No
BERT-base-multilingual	0.712	0.734	0.687	0.710	No
HateBERT [26]	0.681	0.652	0.729	0.688	No
Proposed Transformative Model	0.897	0.912	0.871	0.891	Yes (4-14 hours)

This research provides novel insights into platform-specific algorithmic amplification effects that existing literature has not adequately addressed for Indonesian contexts. While Huszár et al. [2] demonstrated algorithmic amplification of political content on Twitter in English, our findings reveal different amplification patterns for Indonesian content across multiple platforms. TikTok's algorithm showed the strongest amplification of identity-based content (2.8x increase in reach), while Threads demonstrated more balanced algorithmic behavior. These platform-specific findings provide actionable insights for content moderation that existing research lacks.

The research introduces culturally informed intervention protocols that significantly outperform generic approaches used in existing systems. While traditional content moderation relies primarily on content removal or warning labels, our approach demonstrates that culturally sensitive engagement can reduce escalation probability by 67% compared to 23% for generic fact-checking interventions. This represents a paradigm shift from reactive content removal to proactive discourse transformation, addressing fundamental limitations of current content moderation approaches.

Unlike academic studies that often operate on limited datasets or controlled environments, this research demonstrates real-world applicability across multiple platforms with diverse user populations. The system's ability to process 2.3 million words of discourse data while maintaining high accuracy and low false positive rates (15%) addresses practical deployment concerns that existing research has not adequately resolved. The integration of computational efficiency with cultural sensitivity represents a significant advance toward deployable solutions for Indonesian digital spaces.

Despite these advances, several limitations remain compared to idealized solutions. The current approach requires platform-specific model training, limiting immediate cross-platform generalization. Cultural context analysis currently relies on expert knowledge integration rather than fully automated cultural understanding. Future research should focus on developing more generalizable cultural embedding methods and exploring deep learning approaches for automatic cultural context extraction. Additionally, the long-term effectiveness of intervention strategies requires longitudinal evaluation beyond the current 12-month study period.

The integration of computational analysis with cultural understanding offers a more comprehensive approach to digital conflict prevention than purely technical or purely social solutions alone. This multi-disciplinary approach addresses critical gaps in existing research while providing practical tools for preventing identity-based conflicts in Indonesian digital discourse, representing a significant contribution to both academic knowledge and real-world conflict prevention capabilities.

5. Conclusions

This research successfully identified and analyzed the transformation patterns of neutral digital discourse into SARA-based debates across Indonesian social media platforms. The comprehensive analysis of 1,247 discussion threads and 28 in-depth interviews revealed four

distinct transformation phases characterized by decreasing lexical diversity (from TTR 0.67 to 0.29), increasing negative sentiment (from 0.12 to -0.73), and network fragmentation (density decline from 0.34 to 0.12). The developed Transformative Discourse Model achieved 89.7% accuracy in predicting escalation events, representing a 26% improvement over existing methods, with early detection capability 4-14 hours before explicit hate speech emergence. Platform-specific analysis demonstrated significant differences in escalation velocity, with TikTok showing fastest transformation (14.2 hours) and Threads the slowest (31.8 hours), while cross-platform contamination effects occurred in 47% of major escalation events.

The synthesis of findings confirms that discourse transformation follows predictable linguistic and social network patterns that can be computationally detected and analyzed. The research objectives were comprehensively addressed through the identification of linguistic markers (semantic coherence decline, identity term frequency increase), algorithmic amplification effects (2.8x faster spread for inflammatory content), platform-specific escalation characteristics (visual content dominance on Instagram, algorithmic discovery effects on TikTok), and successful predictive model development with cultural context integration. The convergence of quantitative computational analysis with qualitative cultural insights validates the hypothesis that neutral discourse transformation into identity-based conflicts can be predicted and potentially prevented through early intervention strategies. The integration of Digital Critical Discourse Analysis with Social Network Analysis and machine learning approaches proved effective in capturing both micro-level linguistic shifts and macro-level social dynamics that drive escalation processes.

The research findings provide significant implications for multiple stakeholders in Indonesian digital governance and social cohesion efforts. For content moderators and platform developers, the early warning indicators and evidence-based intervention protocols offer practical tools for proactive conflict prevention, moving beyond reactive content removal to preventive discourse management. The culturally informed intervention strategies showing 67% success rates versus 23% for generic approaches demonstrate the importance of cultural competence in digital moderation practices. For policymakers, the research supports evidence-based regulation of digital spaces aligned with Indonesian National Digital Literacy Strategy priorities, particularly in preventing radicalism and maintaining social cohesion in cyberspace. Academic contributions include the Indonesian Digital Discourse Corpus with temporal annotations, cross-platform comparative framework applicable to other multicultural societies, and theoretical advancement of Digital Critical Discourse Analysis for computational social science applications. The findings also inform digital literacy education programs by identifying specific transformation triggers and cultural factors that educators can address in preventive interventions.

Despite these contributions, several limitations require acknowledgment and future research attention. The current study focuses on publicly accessible content and may not capture private group dynamics where significant polarization occurs. Platform-specific model training limits immediate cross-platform generalization, requiring additional research on universal transformation patterns. Cultural context analysis relies partially on expert knowledge rather than fully automated cultural understanding, suggesting need for developing computational cultural embedding methods. The 12-month study period, while comprehensive, requires longitudinal validation to assess long-term intervention effectiveness and evolving platform dynamics. Future research should explore deep learning approaches for automatic cultural context extraction, investigate private group polarization patterns, develop more generalizable cross-platform models, and conduct longitudinal effectiveness studies of intervention strategies. Additionally, expanding the framework to other multicultural societies beyond Indonesia would enhance global applicability and theoretical generalization of the Transformative Discourse Model.

References

[1] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You. Penguin Press, 2011.

[2] F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, dan M. Hardt, "Algorithmic amplification of politics on Twitter," Proc. Natl. Acad. Sci., vol. 119, no. 1, Jan 2022, doi: 10.1073/pnas.2025334119.

[3] T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, 2019. doi: 10.12987/9780300235029.

[4] danah m. Boyd dan N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," J. Comput. Commun., vol. 13, no. 1, hal. 210–230, Okt 2007, doi: 10.1111/j.1083-6101.2007.00393.x.

[5] Monash University Data & Democracy Research Hub, "Rising Levels of Hate Speech on Social Media During the 2024 Election Campaign," Monash University. Diakses: 10 Mei 2025. [Daring]. Tersedia pada: https://www.monash.edu/indonesia/news/rising-levels-of-hate-speech-on-social-media-during-the-2024-election-campaign

[6] Suhaeri dan K. Aditya, "Polarisasi Opini Di Media Sosial Menjelang Pemilu Tahun 2024 Di Indonesia," J. Kebangs. Ri, vol. 1, no. 1, 2023.

[7] I. Syahputra, W. Fajar Riyanto, F. Dian Pratiwi, dan R. Lusri Virga, "Escaping social media: the end of netizen's political polarization between Islamists and nationalists in Indonesia?," Media Asia, vol. 51, no. 1, hal. 62–80, 2024, doi: 10.1080/01296612.2023.2246726.

[8] T. Davidson, D. Bhattacharya, dan I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," in Proceedings of the Third Workshop on Abusive Language Online, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, hal. 25–35. doi: 10.18653/v1/W19-3504.

[9] Z. Waseem, T. Davidson, D. Warmsley, dan I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," in Proceedings of the First Workshop on Abusive Language Online, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, hal. 78–84. doi: 10.18653/v1/W17-3012.

[10] S. Kumar, W. L. Hamilton, J. Leskovec, dan D. Jurafsky, "Community Interaction and Conflict on the Web," in Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, New York, New York, USA: ACM Press, 2018, hal. 933–943. doi: 10.1145/3178876.3186141.

[11] I. Alfina, R. Mulia, M. I. Fanany, dan Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, Okt 2017, hal. 233–238. doi: 10.1109/ICACSIS.2017.8355039.

[12] P. Fortuna dan S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Comput. Surv., vol. 51, no. 4, hal. 1–30, Jul 2019, doi: 10.1145/3232676.

[13] M. KhosraviNik dan J. W. Unger, "Critical discourse studies and social media: power, resistance and critique in changing media ecologies," in Methods of critical discourse studies, R. Wodak dan M. Meyer, Ed., London: SAGE, 2016.

[14] S. Wasserman dan K. Faust, Social Network Analysis. Cambridge University Press, 1994. doi: 10.1017/CBO9780511815478.

[15] Z. Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press, 2017.

[16] N. Fairclough, Critical Discourse Analysis. Routledge, 2013. doi: 10.4324/9781315834368.

[17] K. Charmaz, Constructing grounded theory, 2 ed. London: SAGE Publications, 2014.

[18] M. Rachimoellah, P. H. Lubis, dan N. J. Utimadini, "Digital Activism and Political Change: Challenges of Social Media's Impact on Political Development," J. Middle East Islam. Stud., vol. 11, no. 2, hal. 55–72, Okt 2024, doi: 10.7454/meis.v11i2.177.

[19] R. A. Yanuartha dan L. K. Alfirdaus, "Analisis Wacana Akun Facebook Humor Politik Terkait Pilkada Dki Jakarta Tahun 2017," Cakrawala J. Penelit. Sos., hal. 25–50, 2020.

[20] H. Sazali, U. A. Rahim, R. Farady Marta, dan A. R. Gatcho, "Mapping Hate Speech about Religion and State on Social Media in Indonesia," Commun. J. Ilmu Komun., vol. 6, no. July, hal. 189–208, 2022, doi: 10.15575/cjik.v6i2.

[21] M. Lim, "Many Clicks but Little Sticks: Social Media Activism in Indonesia," J. Contemp. Asia, vol. 43, no. 4, hal. 636–657, Nov 2013, doi: 10.1080/00472336.2013.769386.

[22] J. W. Creswell dan V. L. Plano Clark, Designing and Conducting Mixed Methods Research, Third. SAGE Publications Inc, 2017.

[23] M. Q. Patton, Qualitative Research & Evaluation Methods: Integrating Theory and Practice (4th ed.). SAGE Publications, 2015.

[24] S. P. Borgatti, M. G. Everett, J. C. Johnson, dan F. Agneessens, Analyzing Social Networks, Third. SAGE Publications Ltd, 2024.

[25] aline shakti Franzke, A. Bechmann, M. Zimmer, dan C. M. Ess, "Internet Research : Ethical Guidelines 3.0," 2020. [Daring]. Tersedia pada: https://aoir.org/reports/ethics3.pdf

[26] T. Caselli, V. Basile, J. Mitrović, dan M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, hal. 17–25. doi: 10.18653/v1/2021.woah-1.3.