Multiplatform Topic Modeling Analysis of Gender-Based Violence in Indonesia Using LDA

Annisa Putri Patricia[1*], Nina Setiyawati[2], Dwi Hosanna Bangkalang[3]

[1,2]Informatics Engineering, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Central Java, Indonesia
[3]Information System, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Central Java, Indonesia
E-mail: [1*]672022246@student.uksw.edu, [2]nina.setiyawati@uksw.edu, [3]dwihosanna.bangkalang@uksw.edu

(Received: 10 Apr 2026, revised: 26 Apr 2026, 14 May 2026, accepted: 18 May 2026)

Abstract

Gender-based violence remains a critical social issue in Indonesia, generating substantial public discourse across digital platforms. This study applies Latent Dirichlet Allocation (LDA)-based topic modeling to identify and compare the dominant themes emerging in online discussions of gender-based violence across three platforms: X (formerly Twitter), YouTube, and the news portal Detik.com. A total of 21,000 data points were collected through keyword-based web scraping covering the period January 2020 to October 2025. The two-phase modeling process yielded coherence scores of 0.38 and 0.53 for X, 0.77 and 0.73 for YouTube, and 0.8 and 0.52 for Detik.com across the first and second phases, respectively, reflecting differences in language register and content structure across platforms. A two-phase modeling approach was employed for each platform to progressively refine topic quality and eliminate irrelevant outputs. The results show that LDA successfully identified platform-specific thematic patterns: X captured emotional and psychological dimensions of gender-based violence, YouTube surfaced legislative and advocacy-oriented discourse centered on the PKS Law, and Detik.com produced case-specific topics grounded in legal proceedings and journalistic reporting. Across all three platforms, domestic violence emerged as the most consistently prominent theme. These findings demonstrate that a multi-platform approach yields a more comprehensive and nuanced picture of public discourse on gender-based violence than any single-source analysis, and that LDA is an effective tool for large-scale thematic extraction from heterogeneous online text data.

Keywords: Gender-based Violence, LDA, Social Media, Web Scraping, Topic Modeling

I. INTRODUCTION

Gender-based violence is a social issue that remains a concern for many countries, including Indonesia. In Indonesia itself, the most common forms of violence include sexual violence, domestic violence, psychological violence, physical violence, and economic violence [1], [2], [3]. Advances in science and technology in recent years have also influenced the reporting of gender-based violence issues and made them a topic of public discussion [4]. Social media provides a space for the public to share opinions and experiences regarding various cases of violence [5]. Social media also plays a crucial role for survivors as a space for mutual support and to foster digital activism against violence [6]. The increasing number of gender-based violence cases recorded annually by the National Commission on Violence Against Women (Komnas Perempuan) highlights the urgency of analyzing the discourse unfolding in digital spaces [7].

In Indonesia, social media platforms such as YouTube and X serve as venues for the public to express opinions and share personal experiences. Thousands of posts regarding gender-based violence on these platforms can be analyzed to identify victims' experiences and public reactions [8]. In addition to social media, news portals also play a vital role in reporting on violence and gauging public reactions to violence-related issues [9]. One such example is the news portal Detik.com, which regularly reports on cases of gender-based violence and provides a space for the public to discuss these issues in the comments section. Therefore, social media and news portals can provide a more diverse picture of how gender-based violence is discussed in society.

The reason for selecting the platforms YouTube, X, and Detik.com is that YouTube provides a space for public discussion through comment sections containing public opinions and societal responses to uploaded content on gender-based violence. Meanwhile, X is a platform widely

used to convey opinions quickly and directly, so that responses on this platform can be viewed in real time [10]. News portals such as Detik.com were chosen because they present news from a different perspective and are written in a journalistic style, which is expected to provide a broader picture of how gender-based violence is discussed in society [11].

The objective of this study is to analyze and identify the main topics emerging in public discussions regarding gender-based violence in Indonesia, as well as to interpret how these issues are discussed on social media platforms. The topic modeling approach was employed in the process of identifying discussion topics because it is effective in processing large volumes of text data from online media [12], [13]. The method used in this topic modeling is Latent Dirichlet Allocation (LDA) because it is considered capable of identifying hidden topics within a collection of documents by analyzing the distribution of words appearing in the text [14], [15]. The LDA method has been widely used to analyze public opinion, media discourse, and social issues across various digital platforms [16], [17], [18]. Previous research has demonstrated that LDA is effective in analyzing news articles and social media, enabling the method to uncover specific patterns within large-scale text [19], [20]. Although it has been widely used in research on public opinion analysis, most previous studies have focused on a single data source. This limitation has the potential to provide an incomplete picture. Unlike previous studies, this study contributes by integrating data from three platforms simultaneously, X, YouTube, and Detik.com, thereby producing a more complete map of gender-based violence discourse in Indonesia's digital public sphere [21].

II. RESEARCH METHODOLOGY

This study uses topic modeling with the Latent Dirichlet Allocation (LDA) approach to identify hidden topics or themes in a collection of texts obtained from YouTube, X, and Detik.com. LDA works by determining the probability of words that share similarities and grouping them into several topics [22]. The overall processing pipeline of this study, from data collection to final analysis, is illustrated in Figure 1.

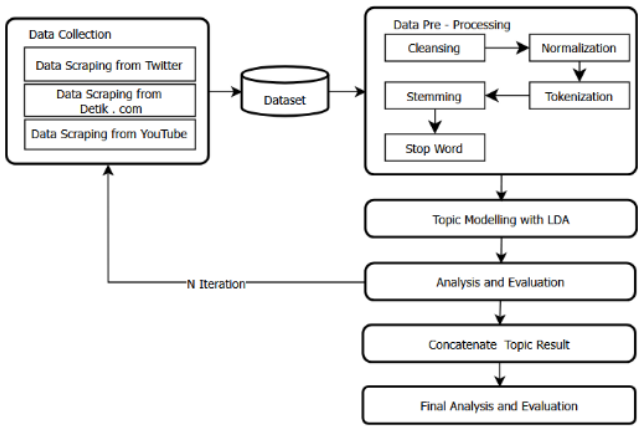


Figure 1. Topic Modeling Processing Pipeline

As illustrated in Figure 1, the processing pipeline consists of the following stages: data extraction via scraping, data preprocessing (which encompasses cleaning, normalization, tokenization, stemming, and stopword removal), topic modeling, model evaluation, combining analysis results, and final analysis. This combination of preprocessing steps was chosen because it is effective in ensuring the interpretability of the analyzed topics [16]. These preprocessing steps are crucial because they determine the quality of the topic modeling results by significantly improving topic coherence [23].

A. Data Collection

The data collection phase was conducted using web scraping methods across three online sources: platform X (formerly Twitter), YouTube, and the news portal Detik.com. Each platform was scraped independently using Python-based scraping tools. For platform X, data were collected from public posts (tweets) using keyword-based queries targeting discussions related to gender-based violence. For YouTube, data were collected from the comment sections of publicly accessible videos relevant to the topic. For Detik.com, data were collected from the comment sections of news articles published under categories related to gender-based violence cases in Indonesia. Data collection was not restricted to official accounts or verified sources; rather, it captured general public posts and comments to reflect broad societal discourse. The keywords used for data collection included domestic violence (kekerasan dalam rumah tangga), physical violence (kekerasan fisik), economic violence (kekerasan ekonomi), sexual violence (kekerasan seksual), and psychological violence (kekerasan psikologis), applied consistently across all three platforms to ensure comparability. Data were collected over a five-year period from January 2020 to October 2025, yielding a total of 21,000 data points distributed across the three platforms. This time range was selected to capture discourse both before and after key legislative milestones, including the passage of the Sexual Violence Crime Law (UU TPKS) in 2022.

B. Preprocessing

The data preprocessing stage is performed to achieve optimal topic modeling by standardizing text structure and cleaning the data. Additionally, raw data is typically unstructured, making it difficult for machines to identify hidden themes; therefore, the data must undergo preprocessing to prepare it for further processing [24]. The steps used in preprocessing are:

a. Cleansing
   This step removes data with a word count of <=1. This is done because text data consisting of short sentences containing only one word likely lacks meaningful context.
b. Normalization
   This step is performed to convert sentence formats into a uniform format so that sentences are easily understood by the model. Normalization steps include case folding, removing URLs, mentions, hashtags, punctuation, and numbers.

DOI: 10.34148/teknika.v15i2.1478

c. Tokenization
In this stage, long sentences are broken down into smaller units called tokens. In this topic modeling process, tokens are used to identify words that are truly significant to the topic and to distinguish them from filler words.
d. Stemming
Stemming is performed to convert words with affixes into their base forms (root words) to standardize the meaning of those words, thereby making word distribution more accurate to support modeling.
e. Stopword Removal
Stopword removal is performed to eliminate words that are frequently used but lack meaningful significance in sentences, thereby allowing the model to more effectively model topics

C. Topic Modeling

Prior to topic modeling, the preprocessed tokens were converted into a numerical representation using the Bag-of-Words (BoW) model. This was implemented using Gensim's corpora.Dictionary utility, which first constructed a vocabulary from all tokens in the dataset. To reduce noise and improve model quality, the dictionary was filtered using filter_extremes with a minimum document frequency threshold of no_below=5 (terms appearing in fewer than 5 documents were removed) and a maximum document frequency threshold of no_above=0.5 (terms appearing in more than 50% of documents were removed). Each preprocessed document was then transformed into a sparse BoW vector using the doc2bow method, resulting in a corpus where each document is represented as a list of (token_id, token_count) tuples. This document-term matrix served as the direct input to the LDA model. The primary model used in this study is Latent Dirichlet Allocation (LDA), a generative probabilistic model that treats each document as a mixture of latent topics, and each topic as a probability distribution over words. LDA assumes that documents are generated by first sampling a topic distribution for each document, then sampling words from the corresponding topic-word distributions. Formally, given a corpus of D documents and a vocabulary of V words, LDA infers K latent topics by estimating two key distributions: θ (theta), the per-document topic distribution, and φ (phi), the per-topic word distribution. Both distributions are governed by Dirichlet priors, α (alpha) and β (beta) respectively, which control the sparsity of topic assignments. LDA was selected as the primary method in this study because the existing literature consistently demonstrates its effectiveness in identifying hidden thematic structures within large-scale text corpora, particularly for analyzing discussion trends and latent topics in social media and news data [14],[19],[20]. Alternative methods such as BERTopic and Non-negative Matrix Factorization (NMF) were considered; however, LDA was preferred due to its probabilistic interpretability, its established track record in Indonesian-language text analysis, and its suitability for the bag-of-words representation of the heterogeneous, multi-platform dataset used in this study. The model was implemented using the LdaMulticore class in Gensim version 4.4.0 with the following parameters: alpha='asymmetric' (allowing each topic to have a different prior probability, which has been shown to perform well on short and informal texts), eta='auto' (learned from data), passes=1, iterations=50, chunksize=2048, and random_state=42 to ensure reproducibility. The number of topics (k) was determined empirically by computing CV coherence scores across a range of k values for each platform and phase, and selecting the k that yielded the highest score. In the context of this study, LDA is applied to extract recurring thematic patterns from comments, tweets, and news articles related to gender-based violence, enabling the identification of the dominant discourse topics and how the public frames and responds to violence across different digital platforms [24].

D. Model Evaluation and Results Analysis

Model evaluation was conducted using the coherence score (CV coherence), a standard metric for assessing the semantic quality of LDA topics. The coherence score measures the degree of semantic similarity between the high-scoring words within each topic, with higher values indicating more interpretable and cohesive topics. In this study, coherence scores were computed across a range of topic numbers (k), and the value of k that yielded the highest coherence score was selected as the optimal number of topics for each platform and processing phase. Coherence scores above 0.6 are generally considered to indicate good topic quality, while scores in the range of 0.3–0.5 reflect moderate interpretability with possible topic overlap [25]. Following topic number selection, the resulting topics were analyzed word by word to identify the dominant theme represented by each topic cluster. Topics were manually evaluated against two explicit exclusion criteria: (1) the dominant words within the topic did not relate to the subject of gender-based violence, and (2) the word distribution did not yield a coherent or interpretable theme. Topic exclusion decisions were made jointly by the researcher and the supervising lecturer to reduce subjectivity and improve the validity of the final topic set. The topics were then visualized using PyLDAvis to inspect inter-topic distances and word distributions, which helped confirm that the selected topics were semantically distinct. In the context of this study, the data processing was performed twice for each platform. The first phase used the full preprocessed dataset to extract an initial set of topics. Topics identified as irrelevant or semantically incoherent in the first phase were then removed, and the remaining data underwent a second round of modeling to produce more refined and focused thematic outputs. The LDA model parameters, including the alpha (document-topic density) and beta (topic-word density) values, were set to their default asymmetric settings as provided by the Gensim library, which have been shown to perform well on short, informal text data [26].

E. Combining Analysis Results

After completing the two processing stages, the topic modeling results from both stages are combined and then analyzed and evaluated during the final analysis stage.

CC BY SA

F. Final Analysis

In this stage, the combined processing results are manually selected and identified to choose topics that are appropriate and relevant to the research focus.

III. RESULT AND DISCUSSION

A. X (Twitter)

Phase 1

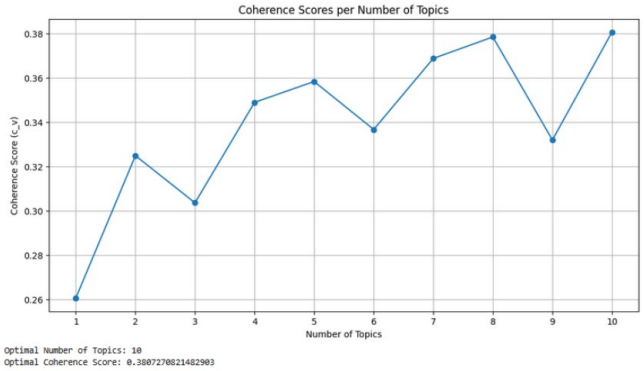


Figure 2. Coherence Score Graph for Phase 1

The first stage of modeling on source X yielded 10 optimal topics (Figure 2) with a coherence score of 0.38, indicating moderate semantic relevance among words within each topic with some degree of overlap possible [27]. The full list of topics is presented in Table 1. The emerging topics were dominated by issues of violence and victim protection; however, topics T#3 and T#7 did not align with the research focus as their word distribution lacked meaningful context and were therefore excluded.

Table 1. Topic Modeling Result of X - Phase 1

Topic	Word
T#1	0.029*"lapor" + 0.018*"kesah" + 0.016*"seksual" + 0.016*"metro" + 0.016*"ulang" + 0.016*"korban" + 0.015*"bunuh" + 0.014*"jaya" + 0.013*"polda" + 0.012*"jakarta"
T#2	0.042*"rumah" + 0.031*"tetangga" + 0.024*"kdrt" + 0.018*"prt" + 0.017*"agama" + 0.016*"seksual" + 0.015*"aneh" + 0.013*"negara" + 0.012*"kerja" + 0.011*"lindung"
T#3	0.025*"akibat" + 0.021*"luka" + 0.018*"orang" + 0.016*"the" + 0.015*"film" + 0.013*"tuntas" + 0.013*"korban" + 0.012*"rusuh" + 0.011*"tubuh" + 0.011*"provinsi"
T#4	0.033*"video" + 0.026*"oknum" + 0.024*"lengkap" + 0.023*"polisi" + 0.022*"duga" + 0.020*"istri" + 0.019*"ajar" + 0.019*"daerah" + 0.019*"berita" + 0.019*"online"
T#5	0.194*"aniaya" + 0.064*"hukum" + 0.032*"laku" + 0.026*"wartawan" + 0.025*"pasal" + 0.021*"duga" + 0.017*"tindak" + 0.011*"bunuh" + 0.014*"orang" + 0.014*"polisi"
T#6	0.080*"perempuan" + 0.070*"gender" + 0.044*"bas" + 0.015*"cepat" + 0.014*"kupang" + 0.014*"seksual" + 0.013*"dpr" + 0.013*"korban" + 0.012*"sorot" + 0.012*"online"
T#7	0.032*"polisi" + 0.026*"pesantren" + 0.021*"santri" + 0.018*"demo" + 0.017*"aksi" + 0.016*"rakyat" + 0.015*"aceh" + 0.014*"ketua" + 0.014*"aparat" + 0.014*"picu"
T#8	0.053*"aniaya" + 0.036*"korban" + 0.027*"leceh" + 0.025*"kdrt" + 0.022*"seksual" + 0.022*"orang" + 0.019*"lapor" + 0.018*"tinggal" + 0.018*"adek" + 0.018*"banget"
T#9	0.107*"anak" + 0.038*"kdrt" + 0.021*"lindung" + 0.023*"umur" + 0.017*"nomor" + 0.016*"republik" + 0.016*"sangka" + 0.014*"undangundang" + 0.012*"hukum" + 0.012*"orang"
T#10	0.057*"seksual" + 0.044*"leceh" + 0.025*"kerja" + 0.019*"aniaya" + 0.017*"korban" + 0.017*"duga" + 0.013*"hukum" + 0.013*"sangka" + 0.011*"negeri" + 0.011*"lapor"

Phase 2

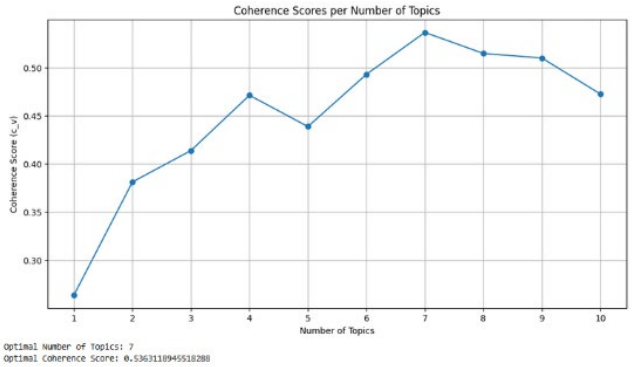


Figure 3. Coherence Score Graph for Phase 2

In the second processing stage, the optimal number of topics was reduced to 7. This reduction indicates a more focused filtering process, producing themes with tighter semantic coherence (Figure 3). As the number of topics decreases, lexical variability within each topic also decreases, allowing the model to group documents into more semantically consistent clusters [28]. The resulting topics are presented in Table 2.

Table 2. Topic Modeling Result of X - Phase 2

Topic	Word
T#1	0.063*"mental" + 0.025*"langsung" + 0.015*"anxiety" + 0.012*"asli" + 0.012*"kumat" + 0.012*"gua" + 0.010*"bagus" + 0.010*"gitu" + 0.008*"ngerasa" + 0.008*"rem"
T#2	0.028*"anxiety" + 0.024*"trauma" + 0.017*"kaya" + 0.015*"gini" + 0.012*"suka" + 0.011*"panik" + 0.009*"anjer" + 0.009*"jam" + 0.008*"kerja" + 0.008*"tahun"

DOI: 10.34148/teknika.v15i2.1478

CC BY SA

Topic	Word
T#3	0.066*"trauma" + 0.029*"orang" + 0.015*"cerita" + 0.010*"sumpah" + 0.009*"takut" + 0.009*"lawan" + 0.009*"sedih" + 0.008*"hidup" + 0.008*"lagi" + 0.007*"suara"
T#4	0.047*"trauma" + 0.032*"mental" + 0.025*"nonton" + 0.010*"bener" + 0.009*"emosi" + 0.009*"liat" + 0.008*"semangat" + 0.008*"tonton" + 0.008*"karena" + 0.007*"semoga"
T#5	0.247*"mental" + 0.019*"gue" + 0.017*"jadi" + 0.014*"liat" + 0.011*"pas" + 0.010*"kali" + 0.010*"trauma" + 0.010*"sekarang" + 0.007*"masa" + 0.007*"susah"
T#6	0.107*"trauma" + 0.010*"kemarin" + 0.009*"pake" + 0.009*"denger" + 0.009*"karena" + 0.008*"paham" + 0.008*"bobol" + 0.008*"inget" + 0.008*"terus" + 0.007*"jadi"
T#7	0.079*"trauma" + 0.023*"gue" + 0.021*"anak" + 0.016*"takut" + 0.013*"sampe" + 0.012*"salah" + 0.012*"sembuh" + 0.011*"coba" + 0.010*"sakit" + 0.009*"emang"

The results of the topic modeling in the second stage revealed differences from the topics discussed in the first stage. In the second stage, the focus of the identified topics shifted from legal aspects to psychological aspects and personal experiences. All seven topics that emerged addressed anxiety disorders and trauma but did not clearly identify the causes of these psychological issues. The topics generated in the first and second stages were then selected and combined to facilitate the final evaluation stage and are listed in Table 3.

Table 3. Interpretation of Combined Results of Phase 1 and Phase 2 - X

Topic	Interpretation
T#1	Proses pelaporan adanya tindak kekerasan seksual ke Polda Metro Jaya
T#2	Terjadinya kasus kekerasan dalam rumah tangga (KDRT)
T#3	Kekerasan dalam rumah tangga yang dilakukan oleh oknum polisi
T#4	Sorotan media terhadap pelaku penganiayaan
T#5	Sorotan media terhadap perempuan sebagai korban kekerasan berbasis gender
T#6	Korban KDRT dan pelecehan seksual yang melapor
T#7	Undang-undang perlindungan anak korban KDRT
T#8	Pekerja diduga alami pelecehan seksual dan penganiayaan
T#9	Gangguan kecemasan yang muncul sewaktu-waktu
T#10	Gangguan kecemasan dan ketakutan sesaat
T#11	Gangguan kecemasan dan ketakutan untuk bercerita
T#12	Trauma akibat menonton sesuatu
T#13	Kesulitan untuk bangkit dari kondisi mental tertentu
T#14	Trauma yang muncul ketika mengingat suatu kejadian tertentu
T#15	Rasa bersalah dan ketakutan

After combining the data, the results show that the topics frequently discussed in the discourse on gender-based violence on platform X pertain to domestic violence—including physical abuse and sexual violence—victims' reporting of violence, and the psychological effects on survivors. This pattern reflects the nature of X as a real-time, opinion-driven platform where users tend to share immediate emotional reactions and personal narratives. The conversational and informal language typical of X posts allowed the LDA model to capture psychological and experiential dimensions of gender-based violence, including expressions of trauma, anxiety, and calls for justice. A noteworthy finding beyond the expected platform characteristics is that trauma- and anxiety-related terms (e.g., "trauma," "mental," "anxiety") dominated the second modeling phase more strongly than direct violence-related terms, suggesting that public discourse on X has shifted toward the psychological aftermath of violence rather than the incidents themselves. This implies that survivors and their supporters use X primarily as a space for emotional processing and peer support, which has practical implications for intervention design: digital mental health outreach targeting gender-based violence survivors may find X a more effective channel than formal reporting platforms. However, the topics that emerged regarding anxiety disorders and trauma did not sufficiently explain the underlying causes of these psychological conditions, likely because X users tend to express emotions rather than provide detailed contextual explanations. This suggests that while X is a rich source for capturing public sentiment and emotional responses, it may be less effective in revealing the structural causes of gender-based violence compared to more formal data sources.

B. YouTube

Phase 1

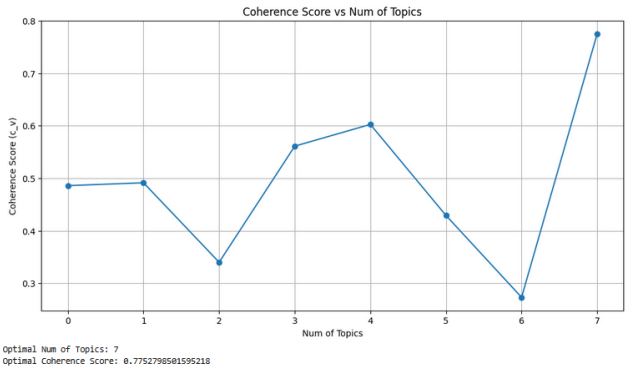


Figure 4. Coherence Score Graph Phase 1

For the YouTube source, Phase 1 yielded 8 optimal topics with a coherence score of 0.77 (Figure 4), indicating semantically cohesive and interrelated topics that are easier to

DOI: 10.34148/teknika.v15i2.1478

interpret. Results are presented in Table 4. Topics T#1, T#5, and T#8 were excluded: T#1 expressed frustration without a specific object, T#5 referred ambiguously to marital relationships, and T#8 focused on religion rather than gender-based violence.

Table 4. Topic Modeling Result of YouTube - Phase 1

Topic	Words
T#1	0.079*"" + 0.039*"banget" + 0.021*"cowo" + 0.018*"malem" + 0.016*"kaya" + 0.014*"cewe" + 0.013*"orang" + 0.013*"gin" + 0.012*"ajar" + 0.011*"bener"
T#2	0.035*"gue" + 0.018*"debat" + 0.015*"perempuan" + 0.012*"udah" + 0.012*"suka" + 0.011*"opini" + 0.010*"bicara" + 0.009*"orang" + 0.009*"malam" + 0.008*"setara"
T#3	0.035*"laki" + 0.034*"perempuan" + 0.017*"langsung" + 0.016*"notif" + 0.015*"gitu" + 0.015*"liat" + 0.013*"nonton" + 0.013*"muka" + 0.013*"equality" + 0.013*"suka"
T#4	0.045*"seksual" + 0.037*"keras" + 0.026*"korban" + 0.022*"gender" + 0.028*"leceh" + 0.018*"tenan" + 0.018*"hukum" + 0.014*"laku" + 0.014*"orang" + 0.013*"tuju"
T#5	0.030*"cowok" + 0.028*"anak" + 0.025*"nikah" + 0.022*"udah" + 0.021*"rumah" + 0.020*"cewek" + 0.017*"perempuan" + 0.015*"keluarga" + 0.013*"temen" + 0.013*"kuliah"
T#6	0.044*"pks" + 0.038*"ruu" + 0.031*"perempuan" + 0.031*"negara" + 0.017*"anggap" + 0.015*"pikir" + 0.015*"atur" + 0.014*"liat" + 0.014*"kayak" + 0.013*"miris"
T#7	0.046*"muda" + 0.037*"anak" + 0.035*"wanita" + 0.024*"mbak" + 0.021*"cerdas" + 0.019*"keren" + 0.017*"dosen" + 0.012*"perintah" + 0.012*"partai" + 0.011*"generasi"
T#8	0.039*"perempuan" + 0.036*"lakilaki" + 0.033*"allah" + 0.025*"manusia" + 0.021*"tuhan" + 0.019*"istri" + 0.014*"hidup" + 0.012*"yesus" + 0.011*"suami" + 0.011*"dalam"

Phase 2

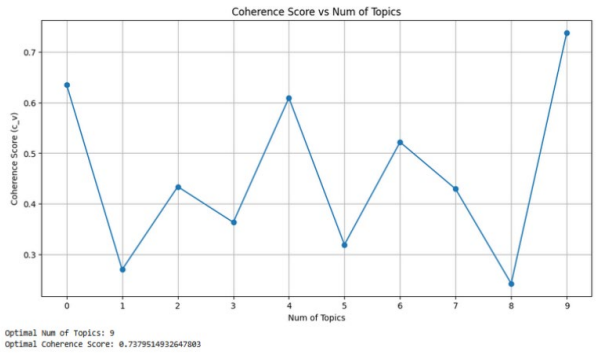


Figure 5. Coherence Score Graph Phase 2

Phase 2 yielded 10 optimal topics with a coherence score of 0.73, consistent with Phase 1 themes (Figure 5). Topics T#2 (podcast host discussion) and T#10 (religious content) were excluded as irrelevant. The combined interpretation of both phases is presented in Table 5.

Table 5. Topic Modeling Result of YouTube - Phase 2

Topic	Words
T#1	0.023*"kdrt" + 0.022*"suami" + 0.020*"perempuan" + 0.018*"rumah" + 0.017*"laki" + 0.014*"istri" + 0.013*"gue" + 0.012*"anak" + 0.010*"cewe" + 0.009*"udah"
T#2	0.131*"" + 0.062*"banget" + 0.025*"keren" + 0.022*"makan" + 0.022*"wanita" + 0.020*"bagus" + 0.020*"grite" + 0.014*"orang" + 0.014*"dok" + 0.013*"cerdas"
T#3	0.026*"dosen" + 0.023*"sakit" + 0.022*"malu" + 0.021*"kampus" + 0.016*"moga" + 0.014*"rasa" + 0.012*"mahasiswa" + 0.012*"ubah" + 0.011*"masa" + 0.011*"kuliah"
T#4	0.036*"liat" + 0.022*"gimana" + 0.018*"bantu" + 0.017*"butuh" + 0.015*"pakai" + 0.015*"bahasa" + 0.014*"korban" + 0.013*"tema" + 0.013*"langsung" + 0.012*"psikiater"
T#5	0.037*"anak" + 0.026*"hidup" + 0.024*"suami" + 0.023*"polisi" + 0.023*"alami" + 0.022*"takut" + 0.021*"lapor" + 0.017*"sakit" + 0.017*"tahan" + 0.016*"cerita"
T#6	0.036*"anak" + 0.030*"negara" + 0.030*"muda" + 0.022*"semua" + 0.017*"edukasi" + 0.015*"sisi" + 0.015*"elvine" + 0.014*"neng" + 0.014*"didik" + 0.014*"maju"
T#7	0.111*"keras" + 0.058*"seksual" + 0.034*"pks" + 0.031*"emosional" + 0.029*"ruu" + 0.025*"korban" + 0.023*"tuju" + 0.016*"orang" + 0.011*"atur" + 0.011*"leceh"
T#8	0.024*"hukum" + 0.018*"lindung" + 0.018*"bicara" + 0.017*"anak" + 0.016*"adil" + 0.016*"laku" + 0.015*"moga" + 0.014*"depresi" + 0.013*"leceh" + 0.012*"gila"
T#9	0.047*"debat" + 0.042*"suka" + 0.036*"anak" + 0.026*"lepas" + 0.023*"muda" + 0.020*"gin" + 0.017*"manusia" + 0.017*"penuh" + 0.016*"pendeta" + 0.016*"tukang"
T#10	0.057*"tuhan" + 0.042*"allah" + 0.031*"manusia" + 0.023*"percaya" + 0.019*"yesus" + 0.015*"dalam" + 0.014*"ajar" + 0.013*"kafir" + 0.013*"hidup" + 0.011*"firman"

The topics that emerged in the first and second stages were then combined and interpreted into sentences. The results of this interpretation are presented in Table 6.

Table 6. Interpretation of Combined Results of Phase 1 and Phase 2 - YouTube

Topic	Interpretation
T#1	Perdebatan tentang isu perempuan
T#2	Isu kesetaraan gender
T#3	Pelecehan seksual dan kekerasan berbasis gender

DOI: 10.34148/teknika.v15i2.1478

Topic	Interpretation
T#4	Keprihatinan terhadap perempuan dan RUU PKS
T#5	Pandangan positif tentang perempuan
T#6	KDRT terhadap istri
T#7	Kesehatan mental mahasiswa di lingkungan kampus
T#8	Dukungan psikologis untuk korban kekerasan
T#9	Tekanan mental karena takut untuk bercerita
T#10	Peran pendidikan dan negara dalam perlindungan anak dan perempuan
T#11	Regulasi UU PKS terhadap korban kekerasan seksual
T#12	Perlindungan hukum dan dampak psikologis korban pelecehan
T#13	Pendeta pelaku KDRT

The topics identified on YouTube are dominated by discussions about the PKS Law (Bill on the Prevention and Eradication of Sexual Violence) and domestic violence. Unlike platform X, where discourse centered on personal emotional expressions, YouTube discussions appear to be more structured and issue-oriented, often tied to specific video content such as documentary coverage, interviews with activists, or educational commentary. This difference suggests that YouTube functions as a platform for more deliberate and informed public engagement with gender-based violence as a societal and legislative issue. The presence of the PKS Law as a prominent topic reflects heightened public awareness and advocacy around legal protections for survivors, indicating that YouTube users are more likely to engage with the policy dimensions of gender-based violence. The overlap in domestic violence topics between YouTube and X indicates a shared public concern regarding this issue across different platform contexts, though the framing and depth of discussion differ considerably between the two platforms.

C. Detik.com

Phase 1

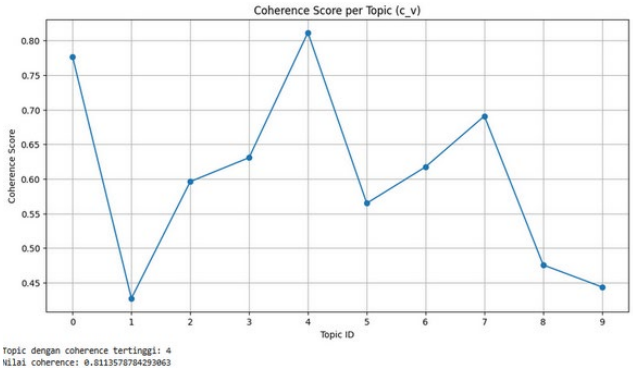


Figure 6. Coherence Score Graph Phase 1

For Detik.com, Phase 1 yielded 5 optimal topics with a coherence score of 0.80 — the highest across all platforms — reflecting the use of standard journalistic vocabulary rather than *slang* or informal terms, which is easier for the model to process (Figure 6). Results are in Table 7. Topics T#1–T#4 showed a clear legal framing, notably centering on the Ferdi Sambo case. Topic T#5 (sports content) was excluded as irrelevant.

Table 7. Topic Modeling Result of Detik.com - Phase 1

Topic	Words
T#1	0.019*"sambo" + 0.017*"putri" + 0.016*"yosua" + 0.014*"hakim" + 0.013*"hukum" + 0.012*"seksual" + 0.011*"jaksa" + 0.010*"ferdy" + 0.010*"bunuh" + 0.009*"dakwa"
T#2	0.012*"anak" + 0.010*"orang" + 0.009*"keras" + 0.008*"seksual" + 0.007*"perempuan" + 0.006*"milik" + 0.006*"didik" + 0.005*"salah" + 0.004*"alami" + 0.004*"sehat"
T#3	0.037*"korban" + 0.023*"laku" + 0.015*"lapor" + 0.012*"polisi" + 0.011*"duga" + 0.010*"seksual" + 0.010*"anak" + 0.008*"orang" + 0.007*"sangka" + 0.006*"aku"
T#4	0.009*"negara" + 0.007*"perintah" + 0.007*"presiden" + 0.007*"kerja" + 0.006*"masyarakat" + 0.005*"menteri" + 0.005*"tingkat" + 0.004*"milik" + 0.004*"usaha" + 0.004*"jalan"
T#5	0.020*"main" + 0.015*"laga" + 0.012*"tim" + 0.011*"poin" + 0.011*"" + 0.010*"menang" + 0.010*"tanding" + 0.010*"grup" + 0.009*"arab" + 0.009*"saudi"

Phase 2

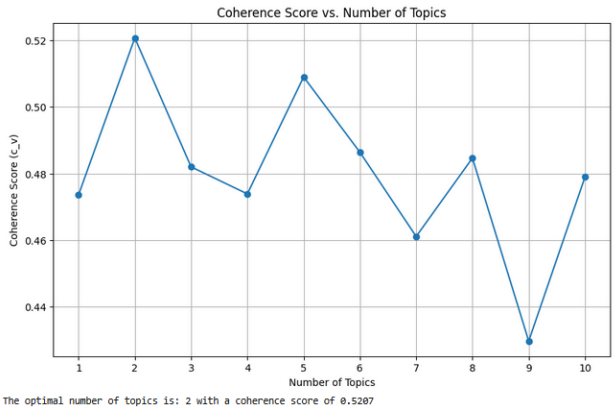


Figure 7. Coherence Score Graph Phase 2

Phase 2 selected 5 topics (second-highest coherence rather than the optimal 2) to avoid overly general themes (Figure 7). Topic T#5 highlighted the Venna Melinda domestic violence case; Topic T#2 was excluded for lacking an interpretable theme. The combined interpretation of both phases is presented in Table 8.

Table 8. Topic Modeling Result of Detik.com — Phase 2

Topic	Words
T#1	0.014*"hukum" + 0.013*"cerai" + 0.012*"sidang" + 0.011*"dakwa" + 0.010*"rumah" + 0.010*"adil" +

DOI: 10.34148/teknika.v15i2.1478

Topic	Words
	0.010*"kdrt" + 0.009*"gugat" + 0.009*"jaksa" + 0.008*"lapor"
T#2	0.007*"rumah" + 0.006*"orang" + 0.005*"lapor" + 0.005*"bandung" + 0.005*"jalan" + 0.004*"qory" + 0.004*"dr" + 0.004*"salah" + 0.004*"keras" + 0.003*"milik"
T#3	0.013*"perempuan" + 0.012*"anak" + 0.011*"keras" + 0.009*"suami" + 0.009*"kdrt" + 0.009*"istri" + 0.006*"istri" + 0.006*"orang" + 0.006*"rumah" + 0.006*"hubung"
T#4	0.028*"korban" + 0.019*"anak" + 0.017*"laku" + 0.016*"istri" + 0.015*"rumah" + 0.014*"polisi" + 0.013*"suami" + 0.012*"lapor" + 0.011*"sangka" + 0.011*"kdrt"
T#5	0.048*"venna" + 0.046*"ferry" + 0.035*"melinda" + 0.030*"irawan" + 0.015*"kdrt" + 0.011*"jatim" + 0.011*"polda" + 0.009*"hotman" + 0.009*"nggak" + 0.008*"periksa"

The results of the first and second modeling phases were then combined, interpreted into a theme, and presented in a Table 9.

Table 9. Interpretation of Combined Results of Phase 1 and Phase 2 - Detik.com

Topic	Interpretation
T#1	Kasus kekerasan Ferdi Sambo dan Istri terhadap Brigadir Yosua
T#2	Anak perempuan yang mengalami kekerasan seksual
T#3	Korban kekerasan seksual laporkan pelaku
T#4	Mengenai perintah presiden terhadap menteri
T#5	Tuntutan hukum pelaku kasus KDRT
T#6	Perempuan dan anak jadi korban KDRT oleh suami
T#7	Anak jadi korban KDRT suami, istri lapor polisi
T#8	Kasus KDRT Venna Melinda oleh Ferry Irawan

The topic of violence that dominates the Detik.com news portal consists of reports on domestic violence cases, the majority of which involve women and children as victims. This differs notably from the results obtained from platform X and YouTube in both topic framing and level of specificity. While X and YouTube captured public emotional responses and advocacy-oriented discussions, Detik.com provided case-specific reporting tied to named individuals and legal proceedings—most prominently, violence cases involving public figures. This distinction reflects the fundamental editorial function of news portals: to document and report on verified incidents that have already entered the legal system.

As a result, the discourse on Detik.com tends to be reactive rather than preventive, focusing on cases that have reached law enforcement rather than emerging social concerns still awaiting institutional attention. In contrast, X and YouTube often surface discussions about cases that have not yet received formal recognition, highlighting the complementary role of social media in filling gaps in mainstream media coverage. Furthermore, the use of standard journalistic vocabulary on Detik.com resulted in higher coherence scores compared to X, where colloquial language and slang introduced greater lexical variability [27]. This finding underscores the importance of using multi-platform data in topic modeling research on sensitive social issues, as each platform contributes a distinct and non-redundant perspective on the same phenomenon.

A cross-platform comparison of the topic modeling results reveals meaningful differences in how gender-based violence is discussed across the three data sources, each reflecting the distinct communicative norms and user behaviors of the respective platform. Platform X produced topics predominantly centered on psychological experiences, emotional responses, and victims' firsthand accounts, consistent with its function as a real-time microblogging platform that prioritizes immediacy and personal expression. YouTube, by contrast, generated topics more closely tied to legislative discourse—particularly the PKS Law—and public education, suggesting that the longer-form nature of YouTube content and its comment sections foster more substantive engagement with the structural dimensions of gender-based violence. Detik.com, as a formal news portal, produced topics anchored to specific documented cases and legal proceedings, reflecting a journalistic framing that emphasizes factual reporting over public deliberation. These platform-specific patterns suggest that the type of discourse captured by LDA is not solely determined by the subject matter, but is also shaped by the affordances and conventions of the medium itself. Consequently, relying on a single platform would risk producing a partial and potentially misleading understanding of how gender-based violence is discussed in the Indonesian digital public sphere. The multi-platform approach adopted in this study therefore provides a more complete and representative picture of public discourse on this issue.

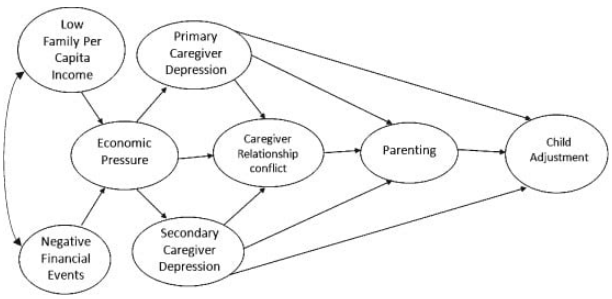


Figure 8. Family Stress Model Theory [29]

The prevalence of domestic violence cases involving psychological and sexual abuse indicates a strong correlation between the two, which often occur simultaneously. Most cases of domestic violence are also triggered by economic problems within the family, such as financial pressure or income instability, which then escalate into physical, psychological, and even sexual abuse. The relationship between economic stress and family violence is conceptually framed by the Family Stress Model (Figure 8), which explains how financial hardship leads to increased family tension and

DOI: 10.34148/teknika.v15i2.1478
CC BY SA

ultimately to acts of violence. This is further supported by the *Family Stress Model*, which states: "Economic uncertainty, such as low income, can lead to increased family financial stress, thereby increasing harsh parenting behaviors and acts of violence against children (Conger et al., 2002)."

In addition to economic factors, the high incidence of domestic violence is influenced by a deeply entrenched patriarchal culture in Indonesian society [30]. The impacts of domestic violence are not merely physical but also include psychological suffering such as trauma, depression, and post-traumatic stress disorder (PTSD), which later disrupt the victims' social lives [31]. Based on the modeling results obtained, the approach used in this study was able to identify and highlight the main themes related to the concept of gender-based violence occurring in Indonesian society. The themes that emerged also align with the cases of violence recorded in the National Commission on Violence Against Women's Annual Report (CATAHU), indicating that the analysis results effectively represent the social conditions prevailing in Indonesia.

Although the model has demonstrated reasonable performance in identifying themes across all three platforms, this study has several limitations that should be acknowledged. First, the data collection process relied on keyword-based scraping, which may have introduced selection bias by excluding relevant discussions that did not use the predefined keywords. Second, the preprocessing stage was insufficient in filtering out all irrelevant and semantically ambiguous words, which contributed to the emergence of topics that were difficult to interpret in some modeling phases. Third, this study did not incorporate temporal analysis, meaning that potential shifts in public discourse over the five-year data collection period (2020–2025) remain unexamined. Future research should address these limitations by refining preprocessing pipelines, applying more rigorous stopword customization for domain-specific vocabularies, and integrating time-series analysis to examine how discourse on gender-based violence evolves in response to policy changes or high-profile cases.

IV. CONCLUSION

This study demonstrates that topic modeling using the LDA method can effectively identify and extract thematic patterns related to gender-based violence from multi-platform online data. A total of 21,000 data points collected from X, YouTube, and Detik.com were analyzed through a two-phase modeling process, with results from both phases combined in the final analysis. The coherence scores across the two phases were 0.38 and 0.53 for X, 0.77 and 0.73 for YouTube, and 0.80 and 0.52 for Detik.com, reflecting the varying linguistic characteristics of each platform. The results show that each platform produced distinct thematic profiles: X captured emotional and psychological responses from survivors and the general public, YouTube highlighted legislative and advocacy discourse particularly around the PKS Law, and Detik.com surfaced case-specific reporting centered on legal proceedings involving public figures. Across all three platforms, domestic violence emerged as the most consistently prominent theme, reflecting its central role in the broader discourse on gender-based violence in Indonesia. Beyond confirming platform-specific communication styles, this study reveals that the type of discourse captured by LDA is actively shaped by the affordances of each medium: X surfaces the psychological burden borne by survivors that often goes undocumented in formal reporting, YouTube amplifies legislative awareness and advocacy that bridges individual experience with policy reform, and Detik.com documents the institutional response to violence that has already entered the legal system. Together, these three layers of discourse form a more complete picture of the gender-based violence landscape in Indonesia than any single platform could provide. These platform-specific differences underscore the value of a multi-source approach, as relying on a single platform would yield an incomplete and potentially skewed picture of how gender-based violence is discussed in the Indonesian digital public sphere. From a practical standpoint, these findings suggest that policymakers and advocates should adopt platform-differentiated communication strategies: leveraging X for mental health support outreach, YouTube for legislative education campaigns, and news portal comment sections for gauging public reactions to legal proceedings. The findings also align with violence statistics documented in the Komnas Perempuan Annual Report (CATAHU), further validating the representativeness of the analysis. Future research should refine preprocessing procedures to address informal language more effectively, explore temporal analysis to capture shifts in discourse over time, incorporate manual validation or inter-rater agreement to strengthen topic interpretation, and consider integrating sentiment analysis to complement the thematic findings produced by LDA.

REFERENCES

[1] Komnas Perempuan, "CATAHU 2024: Menata Data, Menajamkan Arah," Jakarta, Mar. 2025.
[2] Komnas Perempuan, "CATAHU 2023: Momentum Perubahan," Jakarta, Mar. 2024.
[3] Komnas Perempuan, "CATAHU 2022: Kekerasan terhadap Perempuan di Ranah Publik dan Negara," Jakarta, Mar. 2023.
[4] N. Hayati, "Media Sosial dan Kekerasan Berbasis Gender Online Selama Pandemi Covid-19," *Jurnal Hukum, Humaniora, Masyarakat, dan Budaya*, vol. 1, no. 1, pp. 43–52, 2021.
[5] S. Sari, W. Dh, R. Furqan, and F. Hidayatullah, "The Role of Social Media Among Women Survivors of Sexual Violence," *Jurnal Visi Komunikasi*, vol. 24, no. 02, pp. 115–125, 2025.
[6] E. Ratnasari, S. Sumartias, and R. Romli, "Social Media, Digital Activism, and Online Gender-Based Violence in Indonesia," *Nyimak J. Commun.*, vol. 5, no. 1, pp. 97–116, 2021.
[7] Komnas Perempuan, "CATAHU 2025: Menguatkan Data, Mengatasi Kerentanan," Jakarta, Mar. 2026.
[8] I. A. Siregar, F. A. S. N. Haq, "Analisis Sentimen dan

Topik Kekerasan Seksual Perempuan di X 2022-2025 Berbasis Naïve Bayes-LDA," *JOISIE (Journal of Information Systems and Informatics Engineering)*, vol. 9, no. 2, pp. 380–391, 2025.

[9] L. Adhrianti, Yayantika, and N. C. Marlina, "Peran Media Terhadap Fenomena Victim-Blaming dalam Kekerasan Seksual Berbasis Gender," *Jurnal Riset Jurnalistik dan Media Digital*, vol. 4, no. 2, pp. 141–150, 2024.

[10] Musliadi KH, H. Zainuddin, and Y. Wabula, "Twitter Social Media Conversion Topic Trending Analysis Using Latent Dirichlet Allocation Algorithm," *Journal of Applied Engineering and Technological Science*, vol. 4, no. 1, pp. 390–399, 2022.

[11] Z. Munawarah, "Analisis Framing Berita Kekerasan Berbasis Gender Online di Media Liputan6.com dan Antaranews.com," *Medkom: Jurnal Media dan Komunikasi*, vol. 4, no. 2, pp. 139–153, 2024.

[12] R. Hammad, V. C. Hardita, and A. Z. Amrullah, "Topic modeling and sentiment analysis about Mandalika on social media using the latent Dirichlet allocation method," *MATRIX J. Manaj. Teknol. dan Inform.*, vol. 12, no. 3, pp. 109–116, Nov. 2022, doi: 10.31940/matrix.v12i3.109-116.

[13] N. Royani, C. E. Widodo, and B. Warsito, "Topic Modelling Latent Dirichlet Allocation untuk Klasifikasi Komentar pada Layanan Streaming Platform," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 3, Jan. 2024, doi: 10.23887/jstundiksha.v12i3.68492.

[14] A. D. Arko, M. Yusril, H. Setyawan, and R. Andarsyah, "Topic Modeling for Constructing Learning Profiles using Latent Dirichlet Allocation and Coherence Evaluation," *JUTI J. Ilm. Teknol. Inf.*, vol. 23, no. 2, pp. 126–143, 2025, doi: 10.12962/j24068535.v23i1.a1301.

[15] B. B. Wahyujati and F. A. Herawati, "Femicide Framing and Ethics Violation on Indonesian Instagram News Contents," *Jurnal Ilmu Komunikasi*, vol. 22, no. 2, pp. 237–258, 2025.

[16] D. Choirinnisa, F. Alzami, H. Indrayani, A. Rohmani, S. H. Nugraini, R. Zulfiningrumi, *et al.*, "LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 9, no. 2, pp. 849–863, 2025.

[17] B. D. Santosa, N. Fatimah, N. I. Kusumaningtyas, and U. S. Aesyi, "Analisis Kepercayaan Masyarakat Tentang Kepolisian Indonesia di Twitter Menggunakan Latent Dirichlet Allocation (LDA)," *Indonesian Journal on Data Science*, vol. 1, no. 2, pp. 66–76, 2023.

[18] Z. A. Rana, I. T. Haritza, M. A. Fadhillah, and D. A. Purnama, "Analysis of Public Opinion on Disability Services Using Sentiment Analysis and Latent Dirichlet Allocation Topic Modeling Methods," *Jurnal Teknologi Sistem Informasi dan Aplikasi*, vol. 8, no. 3, pp. 151–161, 2025.

[19] M. R. Fahlevvi and Azhari, "Topic Modeling on Online News Portal Using Latent Dirichlet Allocation (LDA)," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 4, pp. 335–344, 2022.

[20] E. S. Negara and D. Triadi, "Topic Modeling Using Latent Dirichlet Allocation (LDA) on Twitter Data with Indonesia Keyword," *Bulletin of Social Informatics Theory and Application*, vol. 5, no. 2, pp. 124–132, 2021.

[21] M. Wetzels, R. Wetzels, E. Schweiger, and D. Grewal, "How Topic Modeling Can Spur Innovation Management," *Journal of Product Innovation Management*, vol. 42, pp. 921–946, 2025.

[22] J. W. Fernanda, "Pemodelan Persepsi Pembelajaran Online Menggunakan Latent Dirichlet Allocation," *Statistika*, vol. 9, no. 2, 2021.

[23] T. Titiana and D. H. Bangkalang, "Analisis Dan Penerapan Topic Modeling Pada Judul Tugas Akhir Mahasiswa Menggunakan Metode Latent Dirichlet Allocation (LDA)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 4, pp. 1275–1287, 2023.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing pada Text Mining Menggunakan Data Tweet 'Mental Health'," *DECODE: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, 2023.

[26] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 481–489, 2024.

[27] H. T. Saputra, A. Damanik, M. M. Shaquille, M. A. Rusydi, M. H. Riziq, and N. S. Zulva, "Analisis Modelling pada Reviews Lazada Indonesia Menggunakan Latent Dirichlet Allocation (LDA) untuk Optimalisasi Strategi Bisnis," *JEKIN (Jurnal Teknik Informatika)*, vol. 5, no. 1, 2025.

[28] Y. Matira, Junaidi, and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi: Journal of Statistics and Its Application*, vol. 4, no. 1, pp. 53–63, 2023.

[29] M. A. Landers-Potts, K. A. S. Wickrama, L. G. Simons, C. Cutrona, F. X. Gibbons, R. L. Simons, *et al.*, "An Extension and Moderational Analysis of the Family Stress Model Focusing on African American Adolescents," *Family Relations: Interdisciplinary Journal of Applied Family Studies*, vol. 64, no. pp. 233–248, 2015.

[30] J. Y. Modiano, "Pengaruh Budaya Patriarki dan Kaitannya dengan Kekerasan dalam Rumah Tangga," *Jurnal Sapientia et Virtus*, vol. 6, no. 2, pp. 129–140, 2021.

[31] M. Janah, P. Pusvitasari, N. Shafira, P. N. Dewi, and N. K. Zahra, "Studi Kasus: Dampak Psikologis terhadap Korban Kekerasan dalam Rumah Tangga (KDRT) dan Implikasinya dalam Proses Peradilan," *Jurnal Ilmiah Penelitian Mandira Cendikia*, vol. 3, no. 4, 2025.

DOI: 10.34148/teknika.v15i2.1478

CC BY SA