



Unifying Knowledge, Reasoning, and Hierarchy for Classifying Harmful Content

Raden Budiarto Hadiprakoso¹, I Gede Maha Putra²

^{1,2} Cryptographic Engineering, Politeknik Siber dan Sandi Negara, Bogor, Indonesia

¹raden.budiarto@poltekssn.ac.id, ²gede.maha@student.poltekssn.ac.id

Abstract

The spread of negative, engagement-driven content online causes significant societal harm, requiring advanced automated moderation tools. However, current classification systems often treat harmful content subtypes as independent, "flat" categories, which hinders their ability to thematically overlap content. This study designed and validated a novel integrated framework to accurately and transparently classify such complex cases. We proposed KG-DToT-HTC, a hybrid framework that synergistically combines three methodologies: a predefined Hierarchical Text Classification (HTC) taxonomy to structure the decision-making process; a domain-specific Knowledge Graph (KG) to provide factual, real-world context; and Decision Tree-of-Thought (DToT) prompting to guide a Large Language Model through an explicit, step-by-step reasoning process. On a real-world dataset of harmful Indonesian news, the proposed framework achieved a state-of-the-art Macro-F1 score of 0.934, representing a nearly 15-percentage point improvement over a zero-shot baseline. Ablation studies confirmed that each component—hierarchy, knowledge, and reasoning—provided a distinct and critical contribution to the final performance. The major conclusion of this study is that a synergistic architecture is essential for the accurate classification of complex harmful content. This work demonstrates a viable path toward "glass-box," interpretable AI moderation systems whose decisions are not only highly accurate but also fully auditable.

Keywords: decision tree-of-thought; harmful content classification; hierarchical text classification; knowledge graphs

How to Cite: R. Budiarto and I Gede Maha Putra, "Unifying Knowledge, Reasoning, and Hierarchy for Classifying Harmful Content", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 1, pp. 133 - 141, Feb. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i1.7115>

Received: August 22, 2026

Accepted: February 6, 2026

Available Online: February 25, 2026

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

News and social media platforms often amplify negative and sensational content to boost user engagement. An additional negative word in a headline increases click-throughs by about 2.3% [1], while negative news is shared on social media nearly twice as often as positive or neutral stories [2]. This has resulted in pervasive exposure to harmful content, with 66% of UK adults reporting encounters with it [3] and 90% of U.S. adults perceiving social media as having negative effects [4].

Beyond attention-seeking negativity, certain types of content—including scams, hate speech, and pornography—cause direct societal harm. Even brief exposure can worsen mental health, with studies reporting heightened anxiety and depressive symptoms after as little as 14 minutes of negative news

consumption [5], [6]. These harms are amplified for children and adolescents, who face elevated risks from predation and cyberbullying, where victims are three times more likely to attempt suicide [7]. At a societal level, misinformation undermines trust in democratic institutions and can fuel unrest [8]. These converging threats create an urgent need for automated, scalable content moderation and public-safety monitoring.

While Transformer models like BERT and RoBERTa have achieved strong results on broad news categories [9-11], they struggle with the nuances of negative-news subdomains like violent extremism and financial scams [12]. Early flat classification approaches, which treat harmful subtypes as independent classes, often fail when themes overlap. For instance, an LSTM-RNN for Indonesian posts struggled when scam and extremist cues co-occurred [13], and a CNN confused "violent

protest” with “violent propaganda” [14]. Even with Transformer-based fine-tuning [15], these models can mislabel a cryptocurrency-funded terror cell as a “scam” instead of “violent extremism,” a high-stakes error with tangible consequences [16], [17]. These limitations demand methods that can encode label structure and external knowledge.

Hierarchical text classification (HTC) offers a partial solution by modeling parent–child relations to enforce top-down consistency [18]. This method organizes labels in a tree, assigning broad categories before refining to subcategories, which improves accuracy and reduces confusion between related labels [19]. The effectiveness of this approach is evident in diverse applications. In multilingual settings, a teacher-student pipeline using GPT-4o and XLM-RoBERTa matched large-model performance on media topics [20]. For low-resource Indonesian toxicity detection, a three-stage hierarchy reduced false positives by 8% compared to a flat model [21]. Recent work has begun integrating external knowledge graphs into hierarchical frameworks to better model semantic dependencies across label levels and improve robustness [22]. However, such integrations remain limited in the context of negative-news classification and public-safety monitoring.

Knowledge-graph (KG)-enhanced classification further improves disambiguation by grounding entities (e.g., “Bitcoin,” “ISIS”) in a network of structured facts. Fusing KG embeddings with text helps separate overlapping labels and propagate external signals like source reliability [23]. This technique has yielded significant gains across various domains, improving precision in cryptocurrency-fraud detection by 7% [24], reducing false positives in fake-news classification by 12% [25], and increasing F1 scores for hate speech by 5% [26]. It has also proven effective in low-resource settings, achieving 86.16% accuracy in Indonesian fact-checking [27]. Yet, the integration of KGs into hierarchical pipelines for negative-news classification remains underexplored.

Even with these structural and knowledge-based enhancements, model decisions can be opaque. Decision Tree-of-Thought (DToT) prompting addresses this by guiding a large language model through a structured reasoning process [28] an evolution of Tree-of-Thought methods that explore multiple reasoning paths [29], [30]. With DToT, the model first forms a coarse rationale (e.g., “This post may be hateful”) and then refines it with targeted sub-questions (e.g., “Does the text contain slurs?”). This approach improves toxic-content classification by up to 4.6% over standard Chain-of-Thought and produces transparent, audit-ready rationales [31]. To date, however, DToT has not been combined with hierarchical taxonomies or KGs.

Despite individual progress with HTC, KG enhancement, and DToT, critical gaps persist. To our knowledge, no prior work unifies these three paradigms

into a comprehensive framework for classifying negative news. Annotated, multi-level datasets for harmful content remain scarce, particularly for low-resource languages [21]. Consequently, there is no integrated method that routes an article through a KG and then traverses a classification hierarchy with DToT-guided reasoning, leaving boundary cases unresolved and explanations fragmented.

To address these challenges, we propose KG-DToT-HTC, a framework that links news entities to a knowledge graph, encodes a hierarchical category tree, and employs DToT to generate stepwise rationales at each level. Recent advances in automated knowledge graph construction using large language models enable dynamic schema induction and scalable triple extraction from web-scale corpora, making such integration increasingly feasible in real-world monitoring systems [32]. Our aims are to (a) reduce misclassification among related harmful topics, (b) enhance interpretability through transparent, level-by-level explanations, and (c) improve generalization in low-resource and overlapping-label scenarios.

2. Methods

This chapter describes the methodology used to develop and evaluate the proposed KG-DToT-HTC framework for classifying harmful Indonesian online content. We begin by presenting an overview of the end-to-end architecture and the role of the gatekeeper module as an initial filtering mechanism (Section 2.1). We then detail the data sources, preprocessing procedures, and train–validation–test split strategy (Section 2.2). Next, we explain the design of the IndoBERT-based Gatekeeper Classifier (Section 2.3) and the two-level hierarchical taxonomy that structures the label space (Section 2.4). The construction of the domain-specific Knowledge Graph for contextual grounding is described in Section 2.5, followed by the Decision Tree-of-Thought (DToT) reasoning mechanism that guides hierarchical label traversal (Section 2.6). Finally, we present the implementation details (Section 2.7) and the experimental setup, including evaluation metrics and ablation configurations (Section 2.8).

2.1 Overview of the KG-DToT-HTC Framework

We design a two-stage pipeline framework to accurately classify harmful online content. A specially trained IndoBERT model first acts as a gatekeeper, performing a rapid initial screening to flag potentially harmful articles. Content identified as harmful is then forwarded to the full KG-DToT-HTC pipeline for detailed hierarchical categorization.

This specialist pipeline integrates three complementary components. The Knowledge Graph (KG) links entities extracted from the article to a compact, domain-specific graph of factual relationships. This grounding provides embedding representations and concise natural-language context that help disambiguate overlapping topics, such as distinguishing cryptocurrency-related

fraud from violent extremism. The Decision Tree-of-Thought (DTotT) mechanism guides the LLM at each hierarchy node to generate multiple candidate labels with supporting rationales, critically compare these alternatives, assign confidence scores, and select the most appropriate category, thereby producing transparent and auditable explanations. Finally, Hierarchical Text Classification (HTC) organizes the label space through a predefined two-level taxonomy, where broad parent categories guide finer-grained subcategories. This top-down traversal constrains the reasoning path and reduces misclassification among nested or semantically related subtypes.

As illustrated in Figure 1, the pipeline begins with the Gatekeeper Classifier. If an article is flagged as harmful, it proceeds to entity linking against the Knowledge Graph before entering the hierarchical DTotT classification stages for Level 1 and Level 2. The final output is a leaf-node label (e.g., “Financial Crimes → Scams & Fraudulent Schemes”) accompanied by a complete chain of human-readable rationales.

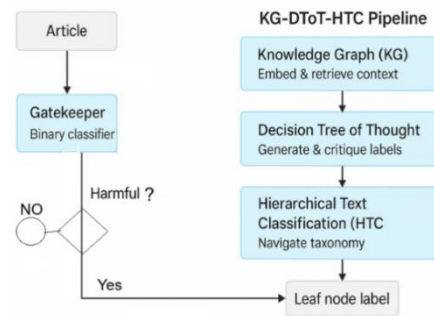


Figure 1. KG-DToT-HTC Framework Pipeline

2.2 Data Acquisition and Preprocessing

We obtained 7,365 monitoring records from the Directorate of Security Operations and Information Control at Indonesia’s National Cyber and Crypto Agency (BSSN) covering October 1 to December 31, 2024. After removing duplicates and records missing critical text fields, a final dataset of 7,140 records was used. We illustrate the dataset with anonymized instances showing the exact fields used by our pipeline in Table 1.

Table 1. Dataset Examples

Title	Description	Type	Cluster	Sub-cluster	Action
“Promo judi...”	“Tautan mengarah ke situs taruhan...”	post	Promotion of Illicit Activities Substances	Online Gambling	block
“JI gunakan Bitcoin...”	“Penggalangan dana anonim...”	news	Ideological Threats & Extremism	Violent Extremism & Terrorism	report
“Kiloan Sabu disita...”	“Polisi menyita paket...”	news	Promotion of Illicit Activities/Substances	Drugs-Related Content	block
“Konten video viral...”	“Video eksplisit...”	tweet	Exploitation, Abuse & Endangerment	Pornography & Sexual Content	block

The subcluster distribution is heavily skewed toward certain types of harmful content. At the top end, Online Gambling (988), Drugs Related Content (945), Pornography (898) and Group (846) together account for over half of all records, reflecting the prevalence of these topics in BSSN’s monitoring logs. The rarest labels—Hate Speech Governments (362), Scam (300),

and Violent Extremism (297)—highlight subdomains with comparatively fewer incidents but potentially higher risk per occurrence. This uneven distribution underscores the need for careful handling of both high volume and low volume classes in our classification pipeline. The distribution of subcluster labels in the dataset can be seen in Figure 2.

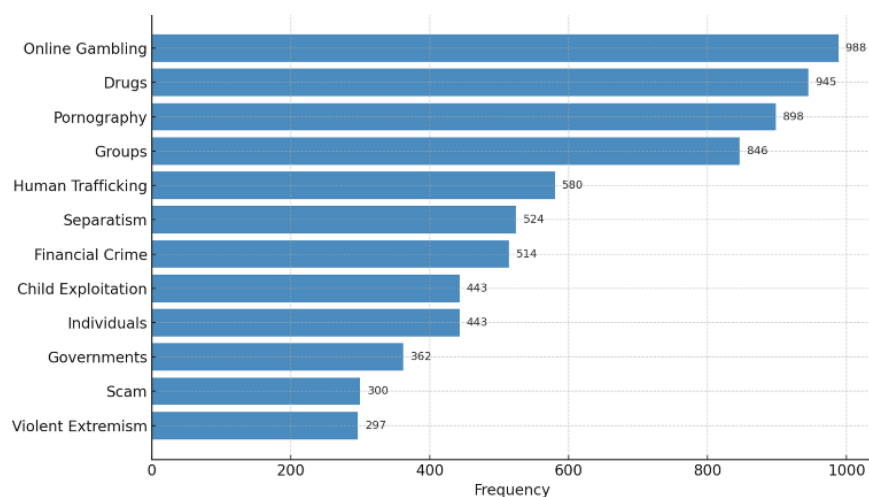


Figure 2. Distribution of subcluster labels

Our preprocessing pipeline concatenates the title and description fields, applies standard text cleaning (lowercasing, punctuation removal), performs stemming using the Sastrawi library, and normalizes common slang using custom dictionaries.

Finally, we perform a stratified split based on the combined cluster-subcluster labels into 70% training (5,000 records), 10% validation (714 records), and 20% testing (1,426 records). All random seeds are fixed (seed=42) for reproducibility. The training and validation sets are used to fine-tune the Gatekeeper Classifier and build the Knowledge Graph. The core DToT-HTC pipeline utilizes a pre-trained LLM without further fine-tuning, ensuring a zero-shot reasoning approach.

2.3. Gatekeeper Classifier Design

To optimize the efficiency of our framework, we employ a binary Gatekeeper Classifier designed to filter out clearly benign content before it reaches the more resource-intensive LLM pipeline. For this task, we fine-tune IndoBERT BaTok, a state-of-the-art BERT model developed for the Indonesian language.

The model was trained on a binary text classification task where we mapped the action field from the dataset to create ground-truth labels: records with an action of “none” were labeled as Benign (0), while those marked for “block” or “report to authorities” were labeled as Harmful (1). The fine-tuning process was conducted on the 5,000-record training set and tuned using the 714-record validation set.

2.4 Taxonomy & Hierarchy Design

To capture the nuanced landscape of harmful content, we designed a two-level hierarchical taxonomy based on BSSN's expert labels. At the root is *Negative Online Content*, which splits into five Level 1 parent categories. Each branch is then refined into specific Level 2 subcategories, as detailed in Table 2.

Table 2. Taxonomy hierarchy

Level 1 (Parent Category)	Level 2 (Child Subcategories)
Financial Crimes & Deception	- Scams & Fraudulent
Exploitation, Abuse & Endangerment	- Financial Crimes
	- Child Exploitation & Endangerment
	- Pornography & Sexual Content
	- Human Trafficking
Ideological Threats & Extremism	- Violent Extremism & Terrorism
	- Separatism & Subversion
Promotion of Illicit Activities/Substances	- Drugs-Related Content
Hate Speech	- Online Gambling
	- Individuals
	- Groups (ethnicity, religion)
	- Governments

2.5 Knowledge Graph Construction

To ground our model in real-world context, we built a compact, domain-specific Knowledge Graph (KG). We started by extracting unique entities from our training data and integrated BSSN's local lexicons. A team of

domain experts then manually curated approximately 500 triples to define the core ontological relationships. These curated triples formed the ontological backbone, which was then expanded by automatically mapping the unique entities extracted from the training data to their respective categories (e.g., (“Sabu”, is a, “Narcotic”), resulting in the final graph of 4,302 triples. At inference time, entities detected in an article are matched to their KG nodes to provide context for the DToT prompt.

The final compact KG serves as a rapid look-up tool during the classification process. When a new article is analyzed, entities within its text are identified and matched to corresponding nodes in the KG. The factual context retrieved from the KG (e.g., embeddings and natural-language facts) is then summarized and injected directly into the DToT prompt. This provides the LLM with the explicit, real-world knowledge needed to differentiate between semantically similar but critically different categories, such as distinguishing a financial scam from terrorism financing.

To assess coverage, we measured the fraction of test-set entity mentions successfully linked to KG nodes (top-1 lexical match with morphological normalization): 91.4% at Level 1 clusters and 84.7% at Level 2 subclusters. We expose the top-k KG facts injected into prompts to enable auditability and error analysis.

2.6 Decision Tree-of-Thought (DToT) Prompting

The heart of our framework is DToT prompting, which The core reasoning component of our framework is Decision Tree-of-Thought (DToT) prompting, which induces structured, comparative deliberation at each node of the label hierarchy, rather than relying on standard zero-shot prediction or a single linear chain-of-thought. For each decision point, the model receives a single, structured instruction that includes the article title and description, a concise summary of relevant facts retrieved from the Knowledge Graph, the current position in the hierarchy, and the set of candidate child categories available at that node.

Given this input, the model first proposes up to three plausible candidate categories and provides a brief rationale for each candidate, explicitly grounded in cues from the text and the injected KG context. The model then performs a comparative self-evaluation by examining the competing rationales, selecting the strongest category, and reporting a confidence score on a 1–100 scale together with a short justification for why the selected option is preferred. This procedure is repeated in a top-down manner: the chosen category determines the next hierarchy node, and the same deliberation process continues until a leaf-level label is reached. As a result, the system outputs both the final classification and a coherent sequence of human-readable rationales along the traversed path, enabling transparent and auditable decisions.

To ensure consistent formatting and downstream parsing, each node-level decision is produced in a constrained JSON structure containing the set of

candidate “thoughts” (category–rationale pairs) and a final “evaluation” field with the selected category, confidence score, and final rationale. Operationally, the prompt content is bounded to maintain stability and cost control: the article snippet is truncated to at most 400 tokens and the KG context summary to at most 60 tokens. When the model’s confidence remains low—specifically, when all proposed candidates receive confidence scores below 60—the system invokes a fallback procedure by expanding the KG context (up to 120 tokens) and re-issuing the decision once. If confidence remains below threshold after this second attempt, the instance is routed to a human reviewer for adjudication.

2.7 Implementation Details

Figure 3 shows a step-by-step “reasoning trace” for classifying a post about Jemaah Islamiyah using Bitcoin. The input snippet (A) is enriched with Knowledge Graph facts (B)—that JI is a terrorist organization and Bitcoin is a cryptocurrency—then a Deliberate Tree-of-Thought module (C) generates competing rationales and provisional labels (e.g., *Financial Crimes*, *Ideological Threats & Extremism*, *Promotion of Illicit Activities*). Following the taxonomy, the system selects the branch *Root* → *Safety Risks* → *Ideological Threats & Extremism* → *L2 Violent Extremism & Terrorism* with confidence 87 (D). A fallback is shown: if the top confidence is below a threshold, the model expands KG context, re-asks once, and escalates to human review if still uncertain.

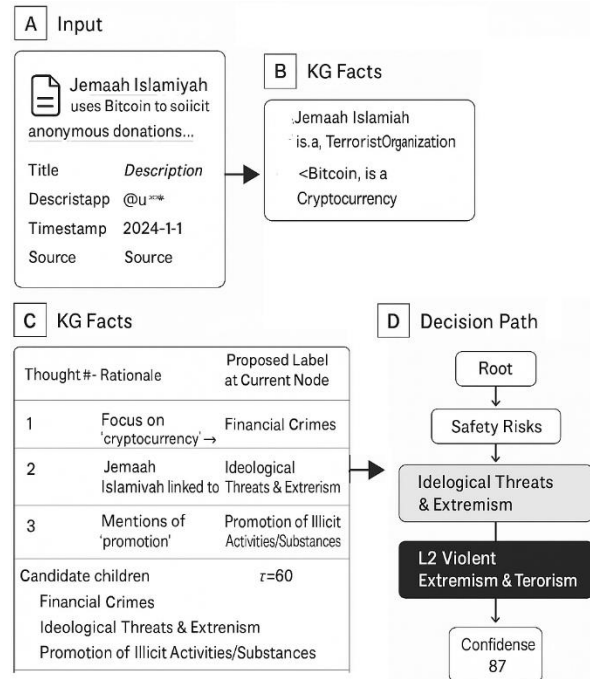


Figure 3. End-to-end decision path

All components were implemented in Python 3.10. We used PyTorch and the Hugging Face Transformers library. The entity linking pipeline was built on spaCy 3.5. Knowledge Graph embeddings were trained using PyTorch-BigGraph. LLM inference was managed via

the OpenRouter API. All training and embedding workloads ran on a single NVIDIA V100 GPU, while LLM inference was handled on a CPU-only instance. Table 3 details the key environmental settings.

Table 3. Environmental Details

Component	Library / Tool	Key Settings
Gatekeeper Classifier (Binary)	Hugging Transformers PyTorch	Face Model: IndoBERT BaTok; lr = 2e-5; batch_size = 16; epochs = 3
Knowledge Graph Embedding	PyTorch-BigGraph	Algorithm: TransE; dim = 128; margin = 1.0; neg_sampling = 1:1
Entity Linking	spaCy 3.5 Custom Patterns	Base model: + id_core_news_sm; no additional fine-tuning
DTot Prompting (LLM Inference)	OpenRouter API	Temperature = 0.0; top_p = 0.9; max_tokens = 256; beam_size = 3

2.8 Experimental Setup & Evaluation Metrics

The experiments are designed in two stages: first, to determine the most effective foundational LLM, and second, to evaluate the contribution of each major component in our framework through targeted ablation studies. Table 4 summarizes the testing scenarios used.

Table 4. Testing Scenarios

Scenario	Configuration	Purpose
Full Comparison	DTot + KG + HTC on GPT 4.1, Gemini 2.5 Pro, Claude Sonnet 4 (identical prompts & hyperparams)	Identify the best base LLM
No-KG Ablation	DTot + HTC (no KG grounding)	Quantify KG's contribution
No-DTot Ablation	Zero-shot HTC + KG (no DTot multi-step reasoning)	Assess DTot's impact
No-HTC Ablation	DTot + KG + flat 12-way subcategory prompt	Evaluate hierarchical routing
Baseline	Zero-shot flat classification (no DTot, KG, or HTC)	Establish a minimal LLM performance baseline

We evaluate model performance across all scenarios using taxonomy-aware and imbalance-robust metrics. Macro F1-score is our primary accuracy measure because it computes an F1-score for each class and then averages these scores equally across all K classes, ensuring that rare but critical categories are not overshadowed by frequent classes such as “Drugs” or “Online Gambling.” Formally, Macro F1 is computed as shown in Equation 1, where K denotes the total number of classes and $F1_k$ denotes the F1-score for class k .

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^K F1_k \quad (1)$$

We also report the Hierarchical F1-score (hF1) to assess how well predictions respect the taxonomy structure (Equation 2). This metric assigns partial credit when the predicted and true labels share ancestors in the hierarchy, penalizing within-branch confusions less

than cross-branch errors. Hierarchical precision (Equation 3) and recall (Equation 4) are computed using the overlap between the ancestor sets of the predicted and gold labels, as defined in. In this equation, N denotes the total number of evaluated instances, y_i is the gold label for instance i , $\hat{y}_i$ is the predicted label for instance i , and $\mathcal{A}(y)$ denotes the set containing label y and all of its ancestor labels in the taxonomy, including the root.

$$hF1 = \frac{2 \cdot hPrec \cdot hRec}{hPrec + hRec} \quad (2)$$

$$hPrec = \frac{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i) \cap \mathcal{A}(y_i)|}{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i)|} \quad (3)$$

$$hRec = \frac{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i) \cap \mathcal{A}(y_i)|}{\sum_{i=1}^N |\mathcal{A}(y_i)|} \quad (4)$$

In addition to accuracy-based metrics, we measure latency, defined as the average response time per LLM call, to assess real-time feasibility in operational content moderation settings. We also measure cost, calculated as the average and total token usage per classification, to estimate the computational and financial implications of each configuration, particularly when relying on commercial LLM APIs.

3. Results and Discussions

This chapter presents the empirical evaluation of the proposed KG-DToT-HTC framework. We first report quantitative results, including foundational LLM comparisons and ablation studies assessing the contributions of KG, DToT, and HTC (Section 3.1). We then provide a qualitative case study to illustrate how the integrated framework resolves complex semantic ambiguity (Section 3.2). Key findings are synthesized to highlight component-specific failure modes (Section 3.3), followed by a comparison with prior work (Section 3.4). Finally, we discuss broader implications, limitations, and directions for future research (Sections 3.5–3.6).

3.1 Quantitative Findings

To identify the most suitable foundational language model, the complete KG-DToT-HTC pipeline was evaluated across three leading LLMs—GPT-4.1, Gemini 2.5 Pro, and Claude Sonnet 4—using identical prompts and configurations. As shown in Table 5, GPT-4.1 achieved the highest scores in both Macro F1 and Hierarchical F1, making it the optimal choice for all subsequent ablation and baseline experiments.

Table 5. Comparative Performance of Foundational LLMs

Model	Macro F1	H-F1	Latency (s)	Tokens
GPT-4.1	0.934	0.931	2.2	1150
Gemini 2.5 Pro	0.923	0.925	2.5	950
Claude Sonnet 4	0.905	0.911	1.4	820

Using GPT-4.1, we conducted ablation studies to isolate the contribution of each major component—KG, DToT, and HTC—and included a zero-shot baseline for comparison. The results presented in Table 6 reveal the

performance drop when key modules are removed. The full model consistently outperforms all ablations, confirming the synergistic value of the integrated design.

Table 6. Ablation and Baseline Scenario Results

Model / Configuration	Macro F1	Hierarchical F1 (hF1)	Avg. Latency (s)	Avg. Cost (tokens)
Full Model (KG-DToT-HTC)	0.934	0.931	2.2	1150
No KG (DToT + HTC)	0.861	0.902	2.1	920
No DToT (KG + HTC, Zero-Shot)	0.884	0.885	1.8	700
No HTC (DToT + KG, Flat)	0.892	0.883	1.9	900
Baseline (Zero-Shot, Flat Only)	0.785	0.790	1.1	650

3.2 Qualitative Analysis: A Case Study in Reasoning

To provide concrete evidence of how the KG-DToT-HTC framework succeeds in untangling nuanced content, we present a qualitative analysis of a challenging example from the test set. The case involves an Indonesian news snippet that intertwines themes of financial technology and terrorism—a combination that proved difficult for our baseline and ablation models.

Test Article: "The Jemaah Islamiyah (JI) group is now using Bitcoin to raise funds anonymously. These funds will be used to finance operations in conflict regions."

Zero-Shot Flat Classifier (Scenario 5): This model predicted "Financial Crimes," fixating on the prominent keyword "Bitcoin" while entirely missing the more severe threat posed by "Jemaah Islamiyah." DToT-HTC without KG (Scenario 2): This model correctly identified the top-level category "Ideological Threats & Extremism" but incorrectly classified the subcategory as "Separatism & Subversion." Lacking specific entity knowledge, it made a plausible but incorrect inference based on the text's tone alone.

Full Framework Success (Scenario 1): The full KG-DToT-HTC pipeline navigated the complexity by leveraging all its components to produce a transparent, evidence-based rationale.

Level 1 Decision: The DToT agent extracted "Jemaah Islamiyah" and queried the Knowledge Graph, retrieving the triple (Jemaah Islamiyah, is_a, TerroristOrganization). This single fact allowed it to confidently discard "Financial Crimes" and select "Ideological Threats & Extremism." Level 2 Decision: Proceeding down the hierarchy, the DToT prompt was now enriched with the context that a terrorist organization was using Bitcoin. This allowed it to precisely differentiate between sibling subcategories and confidently select "Violent Extremism & Terrorism."

3.3 Key Findings

The quantitative results not only confirm the superiority of our integrated framework but also pinpoint the specific weaknesses of simpler approaches. The performance drops in the ablation studies are meaningful—they reveal each component's vital role by showing the specific errors that occur in its absence.

The most significant drop in fine-grained accuracy (Macro-F1: 0.934 $\rightarrow$ 0.861) occurred when the Knowledge Graph was removed. This points to a major failure in understanding context and subtle differences in meaning (semantic disambiguation). Without the KG, the DToT model can still interpret the general tone and structure of a text—correctly identifying a post as an "Ideological Threat"—but it lacks the factual anchor to differentiate between specific real-world entities. For example, it might fail to distinguish a post about a local separatist movement from one about an international terrorist organization because, to the model, they are just names. The KG provides the essential context (is_a: TerroristOrganization) that turns a guess into an evidence-based conclusion.

Removing the DToT reasoning process (Scenario 3) reveals a failure to resolve ambiguity. This model has access to all the necessary facts from the KG and the structure from the hierarchy. However, when faced with conflicting evidence (e.g., an article linking "Bitcoin" to "Jemaah Islamiyah"), it lacks the cognitive mechanism to weigh the facts and make a deliberate judgment. It performs a simple, one-shot inference that can be swayed by strong but misleading signals, leading to a 5-point drop in accuracy. DToT's multi-step "generate-critique-select" process provides the necessary cognitive loop to resolve this conflict effectively.

The "No HTC" model (Scenario 4) highlights a logical consistency failure. While its Macro-F1 score was surprisingly high, it achieved the lowest Hierarchical F1 score. This indicates that while the powerful DToT+KG combination can often find the correct leaf-node label, its errors are more severe and logically inconsistent. It might, for instance, misclassify "Online Gambling" as "Hate Speech" an error that a hierarchical model would never make, as it would have already pruned the "Hate Speech" branch. The hierarchy thus acts as an essential "guardrail," forcing the model's reasoning down a plausible path and preventing these kinds of severe, cross-category mistakes.

3.4 Comparison with Prior Work

Our findings both corroborate and extend earlier research across the framework's three core components. The failure of our flat baseline (Macro-F1 = 0.785) aligns with the struggles of earlier models for Indonesian content [24, 25], confirming that this architectural limitation persists even with modern LLMs. The 7-point Macro-F1 drop in our "No KG" ablation quantitatively mirrors the 5-7% performance gains reported in KG-enhanced classification studies [28, 29]. The 5-point Macro-F1 decline in our "No

DToT" ablation confirms the findings of Zhang et al. [23], who reported a similar ~4.6% accuracy lift using DToT for toxic content detection.

In sum, while prior work has demonstrated the individual benefits of HTC, KG, and DToT, our research provides the first empirical evidence that only their unified application can deliver the accuracy and transparent reasoning needed to fill the gap for a fully integrated harmful-news classification system.

3.5 Broader Implications of Key Findings

Beyond the performance metrics, these findings have significant implications for the design and deployment of AI systems for content moderation.

A Paradigm Shift Towards Interpretable, "Glass-Box" Moderation: The success of the full framework demonstrates a viable path away from opaque, "black-box" classifiers. For sensitive domains like content moderation, where decisions impact free speech and public safety, the ability to produce a transparent, step-by-step rationale is critical for fairness, accountability, and building user trust. Our results show that this interpretability does not have to come at the cost of accuracy.

Validating the "Cognitive Architecture" Approach: The massive 15-point performance leap from the zero-shot baseline to the full framework proves that a powerful LLM alone is not a solution. The success comes from the cognitive architecture, the synergistic combination of a structured hierarchy (long-term memory), a knowledge graph (external memory), and a reasoning engine (the processing unit). This suggests that for complex, real-world problems, a more effective path to improvement is not just scaling up models but designing sophisticated frameworks that enable them to reason and use knowledge effectively.

A Blueprint for Culturally Aware and Context-Specific AI: The framework's reliance on a domain-specific Knowledge Graph provides a tangible blueprint for developing AI systems that are sensitive to local contexts. Instead of depending on generic models trained primarily on Western data, platforms can build and maintain KGs that capture local political entities, cultural nuances, extremist group names, and regional slang relevant to a specific country like Indonesia. This addresses a major flaw in global content moderation efforts, providing a path toward more equitable and effective local deployments.

3.6 Limitations and Future Work

While the proposed framework demonstrates substantial improvements, several limitations remain. First, modality constraints apply because the current system is designed exclusively for textual inputs and therefore cannot capture harm expressed through images, memes, or videos, which are common in real-world online environments. Second, taxonomy rigidity limits expressiveness because the model follows a single hierarchical path, whereas harmful content may

be inherently multi-faceted—for example, hate speech that simultaneously targets a government and an ethnic group. In addition, the fixed taxonomy may not readily accommodate emerging threats, such as novel scam strategies or the promotion of illicit weapons, without manual revision.

This study also raises ethical considerations related to potential bias. The training data, drawn from monitoring records maintained by Indonesia's National Cyber and Crypto Agency (BSSN), may reflect institutional priorities or broader societal biases in determining what is labeled as harmful. Similarly, the manually curated knowledge graph may inadvertently encode subjective judgments from the experts involved. These risks can be partially mitigated by the framework's interpretability: because the system produces transparent, step-by-step rationales, human auditors can scrutinize not only the final label but also the evidence and reasoning that led to it, supporting accountability and bias monitoring.

Future research can address these limitations in several ways. One direction is to extend the framework to a multi-modal setting by incorporating vision–language models so that the DToT agent can jointly reason over textual and visual evidence. Another direction is to increase taxonomic flexibility by supporting multi-label hierarchical classification, enabling an instance to be assigned to multiple relevant categories when appropriate. A further direction is to integrate novelty detection to identify out-of-distribution content; when the model exhibits uniformly low confidence across available categories, the system could flag the instance as potentially harmful but uncategorized and route it to human reviewers, creating a feedback loop for discovering and incorporating new harm categories.

4. Conclusions

This research sets out to address the failure of existing models to accurately classify nuanced and overlapping harmful content at scale. The objective was to design and validate a novel framework that could untangle complex themes by unifying hierarchical classification (HTC), knowledge graph (KG) grounding, and advanced LLM reasoning (DToT). Our proposed KG-DToT-HTC framework has successfully met this objective. The integrated system achieves state-of-the-art performance, with a Macro-F1 score of 0.934 on a challenging real-world dataset. The findings conclusively show that the framework's success is not attributable to any single component, but to their powerful synergy. Ablation studies revealed that removing the KG caused the most significant drop in accuracy, removing DToT crippled the model's ability to resolve ambiguity, and removing the HTC structure resulted in more severe, logically inconsistent errors. Through its step-by-step reasoning process, it enhances model interpretability, moving beyond opaque "black box" systems to provide transparent and auditable rationales for its decisions.

Acknowledgements

This research was funded by a grant from Badan Siber dan Sandi Negara of Indonesia (Grant No. 1491.1/PS/TA.01.02/05/2025)

References

- [1] C. E. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. J. Van Bavel, and S. Feuerriegel, "Negativity drives online news consumption," *Nat Hum Behav*, vol. 7, no. 5, pp. 812–822, Mar. 2023, doi: [10.1038/s41562-023-01538-4](https://doi.org/10.1038/s41562-023-01538-4).
- [2] J. Watson, S. Van Der Linden, M. Watson, and D. Stillwell, "Negative online news articles are shared more to social media," *Sci Rep*, vol. 14, no. 1, p. 21592, Sep. 2024, doi: [10.1038/s41598-024-71263-z](https://doi.org/10.1038/s41598-024-71263-z).
- [3] C. W. Jo, M. Wesołowska, and M. Wojcieszak, "Harmful YouTube Video Detection: A Taxonomy of Online Harm and MLLMs as Alternative Annotators," Nov. 06, 2024, *arXiv:arXiv:2411.05854*, doi: [10.48550/arXiv.2411.05854](https://doi.org/10.48550/arXiv.2411.05854).
- [4] A. Sheth, V. L. Shalin, and U. Kursuncu, "Defining and detecting toxicity on social media: context and knowledge are key," *Neurocomputing*, vol. 490, pp. 312–318, Jun. 2022, doi: [10.1016/j.neucom.2021.11.095](https://doi.org/10.1016/j.neucom.2021.11.095).
- [5] A. Toktarova et al., "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *IJACSA*, vol. 14, no. 5, 2023, doi: [10.14569/IJACSA.2023.0140542](https://doi.org/10.14569/IJACSA.2023.0140542).
- [6] W. M. Johnston and G. C. L. Davey, "The psychological impact of negative TV news bulletins: The catastrophizing of personal worries," *British J of Psychology*, vol. 88, no. 1, pp. 85–91, Feb. 1997, doi: [10.1111/j.2044-8295.1997.tb02622.x](https://doi.org/10.1111/j.2044-8295.1997.tb02622.x).
- [7] A. Schonfeld, D. McNiel, T. Toyoshima, and R. Binder, "Cyberbullying and Adolescent Suicide," *J Am Acad Psychiatry Law*, vol. 51, no. 1, pp. 112–119, Mar. 2023, doi: [10.29158/JAAPL.220078-22](https://doi.org/10.29158/JAAPL.220078-22).
- [8] N. Ayoobi, S. Shahriar, and A. Mukherjee, "Seeing Through AI's Lens: Enhancing Human Skepticism Towards LLM-Generated Fake News," in *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, Poznan Poland: ACM, Sep. 2024, pp. 1–11, doi: [10.1145/3648188.3675136](https://doi.org/10.1145/3648188.3675136).
- [9] S.-Y. Lin, Y.-C. Kung, and F.-Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Information Processing & Management*, vol. 59, no. 2, p. 102872, Mar. 2022, doi: [10.1016/j.ipm.2022.102872](https://doi.org/10.1016/j.ipm.2022.102872).
- [10] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data," *Knowl Inf Syst*, vol. 67, no. 4, pp. 3267–3292, Apr. 2025, doi: [10.1007/s10115-024-02321-1](https://doi.org/10.1007/s10115-024-02321-1).
- [11] X. Men and V. Y. Mariano, "Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19," *IJMecs*, vol. 16, no. 1, pp. 11–22, Feb. 2024, doi: [10.5815/ijmecs.2024.01.02](https://doi.org/10.5815/ijmecs.2024.01.02).
- [12] S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar, "Topic Classification of Online News Articles Using Optimized Machine Learning Models," *Computers*, vol. 12, no. 1, p. 16, Jan. 2023, doi: [10.3390/computers12010016](https://doi.org/10.3390/computers12010016).
- [13] H. B. Aji and E. B. Setiawan, "Detecting Hoax Content on Social Media Using Bi-LSTM and RNN," *bits*, vol. 5, no. 1, Jun. 2023, doi: [10.47065/bits.v5i1.3585](https://doi.org/10.47065/bits.v5i1.3585).
- [14] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, Hangzhou: IEEE, Oct. 2018, pp. 890–893, doi: [10.1109/ITME.2018.00199](https://doi.org/10.1109/ITME.2018.00199).
- [15] J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1098, no. 5, p. 052081, Mar. 2021, doi: [10.1088/1757-899X/1098/5/052081](https://doi.org/10.1088/1757-899X/1098/5/052081).

- [16] S. González-Silot, A. Montoro-Montarros, E. M. Cámara, and J. Gómez-Romero, "Enhancing Disinformation Detection with Explainable AI and Named Entity Replacement," Feb. 07, 2025, *arXiv*: arXiv:2502.04863. doi: [10.48550/arXiv.2502.04863](https://doi.org/10.48550/arXiv.2502.04863).
- [17] J. Chan and Y. Li, "Unveiling disguised toxicity: A novel pre-processing module for enhanced content moderation," *MethodsX*, vol. 12, p. 102668, Jun. 2024, doi: [10.1016/j.mex.2024.102668](https://doi.org/10.1016/j.mex.2024.102668).
- [18] W. Zhao, L. Zhu, M. Wang, X. Zhang, and J. Zhang, "WTL-CNN: a news text classification method of convolutional neural network based on weighted word embedding," *Connection Science*, vol. 34, no. 1, pp. 2291–2312, Dec. 2022, doi: [10.1080/09540091.2022.2117274](https://doi.org/10.1080/09540091.2022.2117274).
- [19] J. Song, F. Wang, and Y. Yang, "Peer-Label Assisted Hierarchical Text Classification," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 3747–3758, doi: [10.18653/v1/2023.acl-long.207](https://doi.org/10.18653/v1/2023.acl-long.207).
- [20] T. Kuzman and N. Ljubešić, "LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification," *IEEE Access*, vol. 13, pp. 35621–35633, 2025, doi: [10.1109/ACCESS.2025.3544814](https://doi.org/10.1109/ACCESS.2025.3544814).
- [21] A. Alamsyah and Y. Sagama, "Empowering Indonesian internet users: An approach to counter online toxicity and enhance digital well-being," *Intelligent Systems with Applications*, vol. 22, p. 200394, Jun. 2024, doi: [10.1016/j.iswa.2024.200394](https://doi.org/10.1016/j.iswa.2024.200394).
- [22] Y. Liu et al., "Enhancing Hierarchical Text Classification through Knowledge Graph Integration," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 5797–5810, doi: [10.18653/v1/2023.findings-acl.358](https://doi.org/10.18653/v1/2023.findings-acl.358).
- [23] Y. Han, A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Knowledge Enhanced Multi-modal Fake News Detection," Aug. 10, 2021, *arXiv*: arXiv:2108.04418. doi: [10.48550/arXiv.2108.04418](https://doi.org/10.48550/arXiv.2108.04418).
- [24] Y. Cao, H. Li, Y. Yu, and S. R. Javaji, "Capybara at the Financial Misinformation Detection Challenge Task: Chain-of-Thought Enhanced Financial Misinformation Detection," in *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, and H.-H. Chen, Eds., Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 321–325. [Online]. Available: <https://aclanthology.org/2025.finnlp-1.38/>
- [25] B. Xie et al., "Multiknowledge and LLM-Inspired Heterogeneous Graph Neural Network for Fake News Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 2, pp. 682–694, Apr. 2025, doi: [10.1109/TCSS.2024.3488191](https://doi.org/10.1109/TCSS.2024.3488191).
- [26] Y. Zhao, J. Zhu, C. Xu, Y. Liu, and X. Li, "Enhancing LLM-based Hatred and Toxicity Detection with Meta-Toxic Knowledge Graph," in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 24747–24760, doi: [10.18653/v1/2025.findings-acl.1270](https://doi.org/10.18653/v1/2025.findings-acl.1270).
- [27] A. P. Muharram and A. Purwarianti, "Enhancing Natural Language Inference Performance with Knowledge Graph for COVID-19 Automated Fact-Checking in Indonesian Language," *J. ICT Res. Appl.*, vol. 19, no. 1, pp. 27–46, Sep. 2025, doi: [10.5614/itbj.ict.res.appl.2025.19.1.2](https://doi.org/10.5614/itbj.ict.res.appl.2025.19.1.2).
- [28] H. Wu et al., "ToTRL: Unlock LLM Tree-of-Thoughts Reasoning Potential through Puzzles Solving," Dec. 26, 2025, *arXiv*: arXiv:2505.12717. doi: [10.48550/arXiv.2505.12717](https://doi.org/10.48550/arXiv.2505.12717).
- [29] S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," Dec. 03, 2023, *arXiv*: arXiv:2305.10601. doi: [10.48550/arXiv.2305.10601](https://doi.org/10.48550/arXiv.2305.10601).
- [30] C. Singh, J. Morris, A. Rush, J. Gao, and Y. Deng, "Tree Prompting: Efficient Task Adaptation without Fine-Tuning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 6253–6267, doi: [10.18653/v1/2023.emnlp-main.384](https://doi.org/10.18653/v1/2023.emnlp-main.384).
- [31] J. Zhang, Q. Wu, Y. Xu, C. Cao, Z. Du, and K. Psounis, "Efficient Toxic Content Detection by Bootstrapping and Distilling Large Language Models," *AAAI*, vol. 38, no. 19, pp. 21779–21787, Mar. 2024, doi: [10.1609/aaai.v38i19.30178](https://doi.org/10.1609/aaai.v38i19.30178).
- [32] J. Bai et al., "AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora," Aug. 01, 2025, *arXiv*: arXiv:2505.23628. doi: [10.48550/arXiv.2505.23628](https://doi.org/10.48550/arXiv.2505.23628).