VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

GRAFAKTA: KNOWLEDGE GRAPH-ENHANCED RAG FOR EXPLAINABLE FACT-VERIFICATION OF INDONESIAN 2024 ELECTION

Fadhil Pratama[1*], Rimba Whidiana Ciptasari[1], Kemas Rahmat Saleh Wiharja[1]

Telkom University, Bandung, West Java, Indonesia[1]
www.telkomuniversity.ac.id[1]
fadhilpratamafp@student.telkomuniversity.ac.id[*], rimbawh@telkomuniversity.ac.id,
bagindokemas@telkomuniversity.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Submitted: 04 June 2026, **Revised:** 07 September 2026, **Accepted:** 09 November 2026, **Published:** 20 November 2026.

Abstract— *This research aims to address the growing misinformation surrounding the 2024 Indonesian presidential election. Surveys show that about 42% of Indonesians are exposed to misleading election-related information. Although some manual fact-checking initiatives have been developed, they are not sufficient to counter misinformation given its vast online presence. Despite the promising results of LLM-based automated fact-checking methods, they may suffer from hallucinations and rely on parametric knowledge. Hence, to address these problems, we present the GraphRAG model, which leverages knowledge graphs to perform explainable fact-checking of Indonesian political claims. To build domain-specific knowledge graphs, we first use 35,965 Indonesian political news articles from Detik.com and CNN Indonesia (2023-2024) and extract 327,793 subject-predicate-object triples using the Llama 3.3 70B LLM. Next, we store them in a Neo4j graph database. We perform semantic retrieval using a multilingual embedding model, BGE-M3, supplemented by multi-hop graph traversal and temporal relevance evaluation. Finally, the LLM generates a structured claim verdict (FACT, HOAX, or UNSUPPORTED) based only on knowledge graph evidence. Evaluation was conducted on a dataset of 200 manually annotated political claims obtained from the main Indonesian fact-checking websites. Vanilla LLMs and standard vector-based RAG were chosen as the primary baselines to demonstrate the incremental value of external knowledge retrieval and structured multi-hop reasoning, respectively. The evaluation shows that our GraphRAG model outperforms vector-based RAG (65.5%, F1-score = 0.767) and vanilla LLMs (48.0%, F1-score = 0.611), reaching 81.5% accuracy and 0.876 F1-score. Moreover, we obtain interpretable reasoning chains and fact correction results.*

Keywords: *Fact-verification, GraphRAG, Indonesian 2024 election, knowledge graph, misinformation detection.*

INTRODUCTION

The 2024 Indonesian presidential election highlighted how rapidly digital misinformation spread, undermining public trust and the perceived credibility of the electoral process. According to CSIS and SAIL's most recent research, about 42% of Indonesians have experienced disinformation related to the 2024 presidential elections, as reported by citizens across 34 provinces [1]. Additionally, Mafindo reported around 2,330 hoaxes in 2023, out of which almost 55% had political motives, more than twice the number reported during the 2019 election [2]. Therefore, there is a dire need for an effective way to automate the verification of political facts within Indonesia.

While manual fact-checking is highly accurate, its speed cannot keep pace with the rapid spread of misinformation on social media platforms. Fact-checking is reportedly slow in Indonesia compared to the rate of spreading disinformation [3]. As a result, Indonesian news agencies and other civil society organizations have launched a digital fact-checking movement and established CekFakta.com, where participating organizations work together to verify and share information [4].

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

However, these efforts are mostly performed manually and cannot keep up with the volume of claims circulating on social media during elections [3].

At first, most existing studies used classical machine learning techniques coupled with handcrafted feature extraction. For example, a web-based fake information detection system using a Passive-Aggressive classifier and n-gram tokenization reported accuracy greater than 90% on Indonesian-language Twitter data [5]. Meanwhile, conventional machine learning approaches have also been applied to Indonesian election-related misinformation detection. A recent study compared K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Random Forest using TF-IDF features for detecting hoax content related to the 2024 Indonesian election, achieving an accuracy of 86.36% with SVM and K-NN under the evaluated configurations [6]. Although effective, this method requires handcrafted feature extraction, leading to a shallow textual representation and an inability to capture deeper semantic meaning and word relationships.

Because of the drawbacks of classical machine learning techniques, subsequent studies have leveraged deep learning approaches to automate feature extraction. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) were common model architectures to solve the hoax detection task. A study reported that a hybrid CNN-LSTM could surpass individual CNNs and LSTMs when used to detect hoaxes on Indonesian political social media, since it can effectively capture local patterns and sequential dependencies within texts [7]. Nevertheless, the drawback of deep learning models in hoax detection is that they require a large amount of labelled data, substantial computational resources, and difficulty in adapting to novel and emerging hoax patterns.

Transformer-based architectures were introduced to overcome the limitations of deep learning models in feature representation, as they showed better performance on many NLP tasks, including text classification. Models such as BERT, RoBERTa, and IndoBERT have successfully captured bidirectional context and enable deeper semantic understanding of the text. Multiple studies reported significant improvements over previous approaches. For example, accuracy of up to 95% was achieved using IndoBERT to detect misinformation in Indonesian Twitter data [7]. IndoBERT was also employed to detect election-related misinformation, achieving 85%-87% accuracy [8]. Even though transformer-based models are very capable of learning representations and achieving higher accuracy than previous methods, they were trained on static data. This makes the models not very adaptive to new information, thus lacking knowledge of newly evolved information.

Leveraging more sophisticated transformer-based capabilities, Large Language Models (LLMs) such as Llama 3 further improved contextual understanding and text generation. LLMs are capable not only of classification but also of producing explanations that help users understand the facts and verify the claims. However, LLMs exhibit the so-called hallucination, which can generate plausible yet factually incorrect text that is not based on source information.

To overcome this limitation, the Retrieval-Augmented Generation (RAG) framework, first introduced by [9], integrates external knowledge retrieval to generate reliable text. RAG is different from transformer-based models, which use static pre-trained data, in that it allows the model to retrieve relevant information from up-to-date fact-checking repositories or credible online sources, so the output can be more factual. A RAG-based hoax detection system in Indonesia was developed, achieving 79.5% accuracy and demonstrating higher reliability than LLMs alone by presenting evidence for verification [9]. An advanced RAG method, such as RAG-Fusion, could achieve 88% accuracy against various forms of misinformation through a multi-query retrieval approach [10]. Even though the accuracy is not as high as previously trained models, such as IndoBERT (95%), the system based on LLM and RAG is usually tested on more challenging datasets using a zero-shot or out-of-distribution setup and requires generating an explanation instead of binary classification, which means that the system demonstrates a significantly more adaptable system towards new information outside training data than previously fine-tuned models.

A parallel line of research has explored the use of Knowledge Graphs (KGs) as structured external knowledge sources for fact verification. Unlike unstructured text corpora, KGs encode real-world entities and their semantic relationships as triples (subject, predicate, object), enabling more principled, interpretable reasoning over factual claims. The transformer-based models are trained on static, pre-trained data, but RAG gives the models the ability to pull relevant information from current fact-checking repositories or trustworthy online sources so the output can be more factual. More recent research has explored knowledge-path-based approaches for automated fact-checking. Wang et al. introduced Path-FC, a fact-checking

CC BY NC

model that incorporates knowledge paths extracted from external knowledge graphs to enrich claim and evidence representations, demonstrating that knowledge-path retrieval can improve false-claim identification [11]. Similarly, Opsahl investigated simplified subgraph retrieval for fact verification using structured knowledge-graph evidence from the FactKG benchmark, showing that efficient graph-based evidence retrieval can achieve competitive verification performance while reducing computational requirements [12].

Later works have moved towards integrating deep learning and pre-trained language models into KG-based pipelines. Liu et al. (2020) proposed KGAT, a Kernel Graph Attention Network for fact verification that constructs an evidence graph from Wikipedia documents and uses kernel-based attention to extract the most claim-relevant evidence before verification, achieving significant improvements over text-only baselines on the FEVER benchmark. Similarly, Mayank et al. (2022) proposed DEAP-FAKED, a framework that jointly encodes news article content using NLP and embeds KG entities through tensor decomposition [13]. The authors demonstrated that the complementary use of structured knowledge and textual features improves fake news detection in political and healthcare domains.

More recent work has presented benchmark datasets and sophisticated reasoning mechanisms for KG-based fact verification. Kim et al. (2023) proposed FactKG, a large-scale dataset containing 108,000 claims from DBpedia, annotated with five types of reasoning, including one-hop, conjunction, existence, multi-hop, and negation, providing a systematic evaluation framework for structured fact verification [14]. INSPECTOR is a graph attention network model proposed by Liu et al. (2022) for pair-wise fact checking that detects semantic inconsistencies between claim pairs. The work of shows the utility of KGs for comparative reasoning beyond single-claim verification [15]. Nikopensius et al. (2023) proposed a reinforcement learning-based method to further promote the explainability dimension, in which an RL agent traverses KG paths to prove or disprove a claim, and the resulting paths are human-readable explanations [16]. The design prioritizes transparency and trust, besides classification performance.

Taken together, these findings underscore a number of advantages of KG-based approaches: They provide structured, traceable chains of reasoning; they are less prone to hallucination than purely parametric models; and they can be combined with LLMs to improve coverage and expressiveness. However, KG-based systems are limited by the completeness and currency of the underlying knowledge base, and they may have difficulty with naturally phrased claims that do not map neatly to structured triple representations. These challenges are particularly pertinent to the Indonesian hoax detection scenario, where domain-specific KGs are still limited, and motivate integration strategies that combine structured knowledge with retrieval-augmented generation.

There are some important problems to be solved for hoax-detection systems to be completely accurate, adaptive and reliable. Feature representation and data dependence are challenges in machine learning and deep learning models. Yet, even with their impressive abilities, LLMs are not immune to adversarial attacks, input manipulation, and the danger of spreading misinformation if their security alignment is compromised [17]. They also have the disadvantages of using old parametric knowledge and a high chance of hallucination [7]. The RAG model alleviates some of these issues by grounding the generated responses into retrieved factual evidence. But there are practical implementation challenges. However, RAG systems are still susceptible to false negatives when faced with subtle hoaxes or limitations in the retrieved context, highlighting the need for improved explainability in the model's reasoning process [9]. Additionally, state-of-the-art RAG frameworks suffer from operational bottlenecks such as large latency spikes in retrieval of complex queries [10].

Table 1. Comparison of Related Studies

Title	Approach	Performance	Limitations
Web-based Fake Information Detection [5]	Classical machine learning: Passive-Aggressive classifier.	>90% accuracy.	Requires handcrafted feature extraction; shallow textual representation; cannot capture deep semantic meaning or word relationships.
Hybrid CNN-LSTM for Indonesian Political Hoax Detection [18]	Hybrid CNN-LSTM combines local pattern capture (CNN) and sequential dependency learning (LSTM).	Outperforms individual CNN and LSTM models.	Requires large labelled datasets; high computational resources; difficulty adapting to novel and emerging hoax patterns.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

Title	Approach	Performance	Limitations
IndoBERT for Misinformation Detection on Twitter [7]	IndoBERT bidirectional transformer pre-trained on Indonesian text.	~95% accuracy.	Trained on static data; not adaptive to new or evolving information; relies on outdated parametric knowledge; hallucination risk.
IndoBERT for Election Misinformation Detection [8]	IndoBERT fine-tuned on election misinformation dataset.	85%–87% accuracy.	Same static knowledge limitation; cannot adapt to post-training misinformation; lacks knowledge of newly evolved information.
LLM-based Hoax Detection with Explanation using Llama 3 [17]	Large Language Model (Llama 3) generative, instruction-following model.	Capable of classification and explanation generation.	Prone to hallucination; vulnerable to adversarial attacks and input manipulation; may propagate false information if security alignment is compromised.
RAG-based Hoax Detection System in Indonesia [9]	Retrieval-Augmented Generation (RAG) integrated with LLM.	79.5% accuracy.	Still produces false negatives on subtle hoaxes; limited explainability in reasoning process; retrieved context limitations reduce reliability.
RAG-Fusion for Misinformation Detection [10]	RAG-Fusion advanced RAG with multiple parallel query generation and result fusion.	88% accuracy across various misinformation forms.	Significant latency spikes during complex query retrieval; operational bottlenecks in production; evidence traceability and transparency of reasoning still lacking.

Source: (Research Results, 2026)

As summarized in Table 1, existing approaches exhibit critical limitations that GraFakta is specifically designed to overcome. First, classical machine learning and deep learning models [5] are highly dependent on manual feature engineering or large labelled datasets and suffer from static representations. However, transformer-based models such as IndoBERT [7], [8] can improve semantic understanding. Still, they are constrained by their static pre-training data, which prevents them from adapting to newly evolved political events without expensive retraining.

GraFakta solves this static knowledge problem by using a dynamically constructible Knowledge Graph (KG) that can be continuously updated with recent news. Second, while generative LLMs [17] provide robust reasoning, they introduce the risk of hallucination and reliance on unverified parametric knowledge. GraFakta explicitly mitigates this by restricting the LLM's reasoning strictly to the retrieved KG evidence. Finally, although standard RAG frameworks [9], [10] mitigate hallucination, they typically rely on flat vector similarity search, which struggles to perform the multi-hop reasoning required to debunk subtle political hoaxes.

Moreover, while recent studies have shown the potential of GraphRAG and KG-based reasoning for fact-checking, current frameworks are insufficient in the specific context of Indonesian political fact-checking for several reasons. First, the major knowledge graphs (e.g., DBpedia, Wikidata) and benchmark datasets are largely English-centric and do not comprehensively cover localised Indonesian political entities, regional terminology, and dynamic electoral events. Second, political mis- and disinformation in Indonesia often depends on subtle temporal nuances, such as the rapidity of party coalitions or changes in official statements in the pre-campaign, campaign, and post-election periods. Standard GraphRAG models typically rely on flat semantic similarity and cannot capture such chronological shifts without explicit temporal reasoning mechanisms. Third, the highly contextual nature of Indonesian political news requires a domain-specific knowledge base to track specialised semantic relations, which general-purpose KGs lack. Thus, relying on existing GraphRAG architectures without domain-specific tuning and temporal awareness is insufficient for verifying complex narratives about the 2024 Indonesian election.

Building on these design principles to overcome the aforementioned limitations, we propose an explanation-oriented fact verification model based on GraphRAG, which uses a specialized knowledge graph constructed from subject-predicate-object triples extracted from the corpus of Indonesian political news. Large language models are used to generate such triples, which are then stored in a Neo4j graph database. Following that, retrieval with embeddings is performed to generate structured decisions, including *FAKTA* (FACT), *HOAKS* (HOAX), and *TIDAK TERBUKTI* (UNSUPPORTED), along with evidential support.

The main novelty of this research is the integration of multi-hop graph traversal, temporal relevance scoring, and contradiction detection, tailored to the dynamic, low-resource context of the 2024 Indonesian presidential election. In particular, our contributions are summarised as follows:

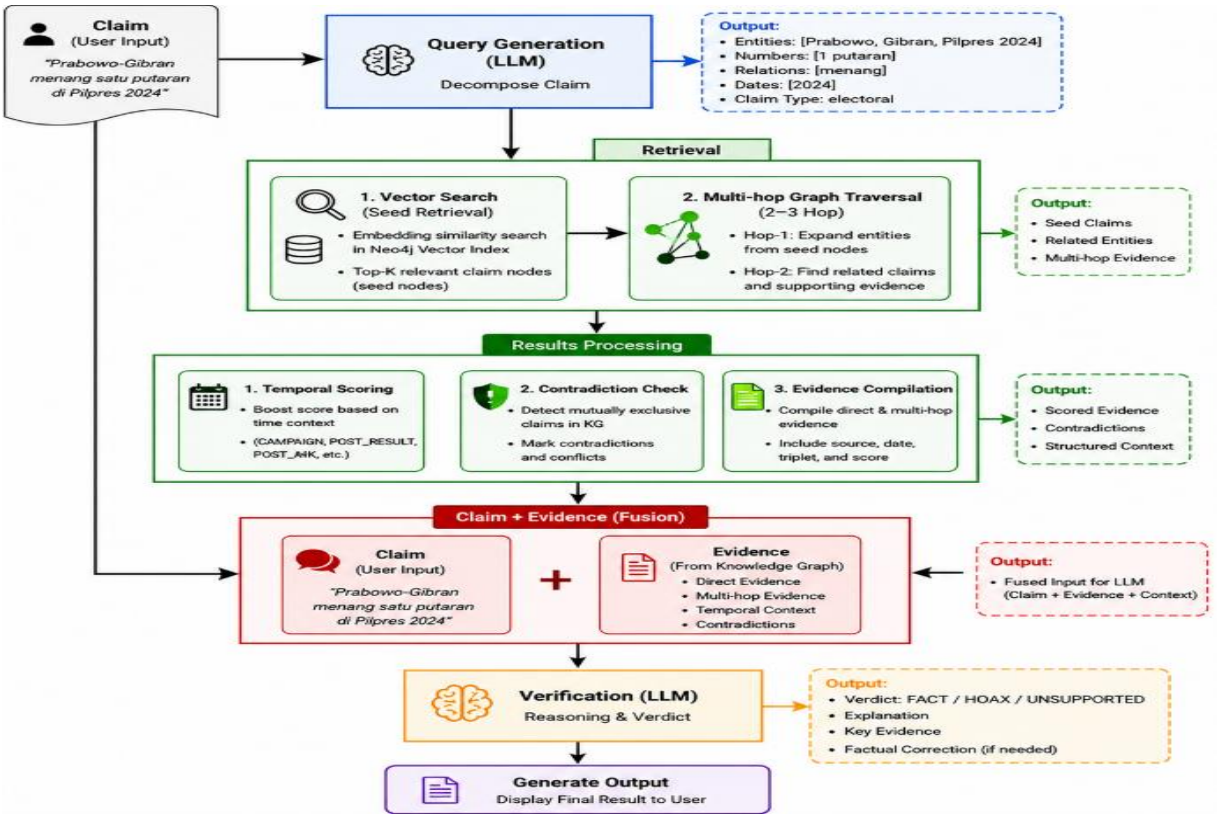
1. Create the first domain-specific knowledge graph for the 2024 Indonesian presidential election, leveraging LLM-assisted extraction to construct a highly contextualised knowledge base from local news articles (2023-2024).

CC BY NC

2. Perform empirical experiments comparing the proposed framework with standard RAG and LLM baselines to quantify the individual and combined contributions of graph-based reasoning to debunking local political misinformation.
3. Publicly release the extracted SPO triple dataset comprising 327,793 triples from Indonesian political news to support reproducibility and future research in low-resource language fact-checking.

MATERIALS AND METHODS

In this work, we present a GraphRAG framework for explainable fact verification in Indonesian political news related to the 2024 presidential election.



Source: (Research Results, 2026)

Figure 1. Proposed Model

The proposed framework consists of three major stages as illustrated in Figure 1: (1) Data Preparation, covering news article collection and knowledge graph construction via LLM-assisted SPO triple extraction; (2) Graph Retrieval, encompassing embedding generation, vector indexing, and multi-hop semantic evidence retrieval from the Neo4j graph database; and (3) LLM Inference, where the compiled graph evidence is used to generate structured verdicts (FACT, HOAX, or UNSUPPORTED) using the Llama 3.3 70B model.

Data Preparation and Preprocessing

Table 2 presents a sample of the collected dataset. Political news articles were systematically collected from two major Indonesian online news platforms: Detik.com (via the dedicated Detik Pemilu section) and CNN Indonesia, for the period from January 2023 to December 2024. The sources were chosen because they cover the 2024 Indonesian Presidential Election extensively and because they are reputable, established media outlets. Web scraping was used to collect data from the political news category. The scraping process resulted in 35,965 news articles that were divided into four dataset partitions, namely 2,477 articles from CNN Indonesia (2023), 2,725 articles from CNN Indonesia (2024), 18,331 articles from Detik.com (2023), and 12,432 articles from Detik.com (2024). Each article was stored alongside its metadata (article title, publication date, full URL and source identifier), which was then used for citation references during verdict generation.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

Table 2. Sample of Collected Indonesian-Language News Articles Dataset

Date	Title	Content	URL
2023-01-01	*Jelang Pemilu 2024, Puan Harap TNI Netral*	*Ketua DPR RI Puan Maharani meminta Panglima TNI Laksamana Yudo Margono memastikan (...)*	*https://news.detik.com/pemilu/d-6492847/jelang-pemilu-2024-puan-harap-tni-netral-dari-politik-praktis*

Source: (Research Results, 2026)

Knowledge Graph Construction

Table 3. Sample of Extracted SPO Triples from Indonesian News Articles

Claim	Subject	Relation	Object
Moeldoko memastikan bantuan sosial beras tetap dilanjutkan.	*Moeldoko*	*MEMASTIKAN*	*Bantuan Sosial Beras*
Program bansos telah dilaksanakan sejak September 2023	*Bantuan Sosial Beras*	*TELAH_DILAKSANAKAN_SEJAK*	*September 2023*
Bansos akan berlanjut di tahun 2024	*Bantuan Sosial Beras*	*AKAN_BERLANJUT_DI*	*Tahun 2024*

Source: (Research Results, 2026)

The collected news articles were processed to generate structured facts in the form of subject-predicate-object triples. For the extraction task, the Llama 3.3 70B Versatile model obtained via the Groq API was chosen for its excellent multilingual instruction-following ability and high inference efficiency, which are well-suited to processing Indonesian texts. The text of each article, together with its title, was fed to the model with a structured extraction instruction indicating that it should find and extract information about the names of entities and their semantic relations present in the input text. Table 3 presents representative examples of SPO triples extracted from the collected news articles, illustrating how entities and their semantic relations are captured from Indonesian political text.

Our system uses an instruction-guided Open Information Extraction (OpenIE) paradigm to extract semantic relations. While the Llama 3.3 70B model is not tied to a strict, pre-defined ontology (which is needed to accommodate the fluid and nuanced lexicon of Indonesian political events), predicate generation is tightly governed by prompt instructions. These prompt constraints force the LLM to extract a concise, fact-based relation that directly reflects the source text, with strict structural formatting (e.g., uppercase formatting and underscore delimiters for multi-word relations), thereby avoiding verbose or hallucinated predicates, as illustrated in Table 3. Any residual lexical variations in predicates (e.g., extraction of *'MENDUKUNG' 'MEMBERIKAN_DUKUNGAN'*) are effectively compensated in the subsequent retrieval phase. The BGE-M3 model encodes the whole natural language representation of the SPO triple, so semantically synonymous relations are clustered together in the dense vector space. This enables robust evidence retrieval despite minor differences in phrasing.

Table 4. Knowledge Graph Dataset Statistics

Source	Year	Articles	Triples Extracted
CNN Indonesia	2023	2,477	14,087
CNN Indonesia	2024	2,725	10,664
Detik.com	2023	18,331	179,165
Detik.com	2024	12,432	123,877
Total	**2023-2024**	**35,965**	**327,793**

Source: (Research Results, 2026)

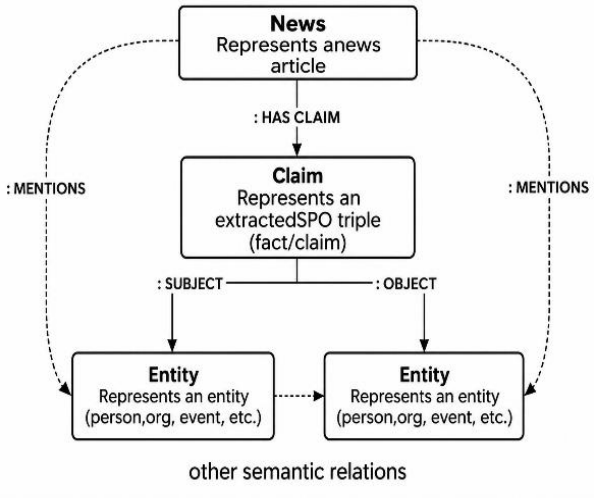
The extraction process was applied systematically across all four dataset partitions. The resulting distribution of articles and extracted triples per source and year is summarized in Table 4. In total, the extraction process produced 327,793 triples across four partitions of the dataset, as detailed in Table 4. The generated triples were then saved in the following format: head entity, relation type, tail entity, source title, publication date, and source URL.

We also performed a manual human evaluation on a random sample of 200 extracted SPO triples and their source articles to assess the reliability of the constructed knowledge graph. The evaluation was conducted by three independent annotators with academic backgrounds in cybersecurity and digital forensics. We applied a majority voting scheme based on three strict criteria: Subject Correctness, Relation Appropriateness, and Object Correctness. A triple was deemed valid if two out of three annotators agreed that the three components accurately reflected the factual information in the source text without semantic hallucination.

The overall extraction accuracy from this manual evaluation was 92.5 percent. The few errors we saw (7.5%) were mostly due to unresolved coreferences in the news text (e.g., extracting pronouns instead of explicit named entities) and extracting overly complex object clauses. This high-accuracy metric indicates that the zero-shot

CC BY NC

extraction capability of the Llama 3.3 70B model is reliable for constructing a domain-specific political knowledge graph.



Source: (Research Results, 2026)

Figure 2. Knowledge Graph Schema

Furthermore, to manage entity name variations across the 327,793 extracted triples, our framework adopts a 'soft resolution' strategy mitigated by dense semantic retrieval rather than explicit graph-level node merging. Due to the open-ended nature of zero-shot LLM extraction, variations of the same entity (e.g., 'Jokowi' versus 'Joko Widodo') may be stored as distinct Entity nodes. However, this structural separation does not degrade retrieval performance because the claim_vector_index relies on BGE-M3 embeddings, which inherently map syntactically varied but semantically identical entities into closely aligned dense vector spaces. Consequently, during the initial vector query, the system simultaneously retrieves these varied Claim nodes as highly ranked seeds. This semantic grouping ensures that all relevant evidence is captured for the multi-hop traversal, making the GraFakta architecture robust against nomenclature variations without the computational overhead of database-level entity deduplication.

Further, the triples were loaded into the Neo4j graph database, where each article is represented by a News node, and Entity nodes are connected to it via the MENTIONS relationship and related to one another based on their semantic connections expressed by the triple relation type. The resulting graph schema, comprising News, Entity, and Claim nodes interconnected through MENTIONS, HAS_CLAIM, SUBJECT, and OBJECT relationships, is illustrated in Figure 2.

Embedding Generation and Vector Indexing

The BGE-M3 encoder [19] is applied locally to encode each SPO triple inserted into Neo4j into a dense vector representation to facilitate semantic retrieval in the knowledge graph. Prior to encoding, we convert each triple into a natural language claim string using the template: "[head] [relation] [tail]" which we find produces more semantically meaningful embeddings for retrieval than raw structured triples. In this current project, we have four main reasons to choose BGE-M3. First, it has native support for over 100 languages, including Bahasa Indonesia. Secondly, it provides a unified framework to perform dense-vector, sparse-vector and multi-vector retrieval tasks in a single model. Third, the model can accept input sequences of a few tokens up to 8,192 tokens, and thus is able to handle short SPO triples as well as longer news article titles. Fourth, BGE-M3 enables cross-lingual semantic matching, which is especially useful as political claims might be expressed differently than in the original news text.

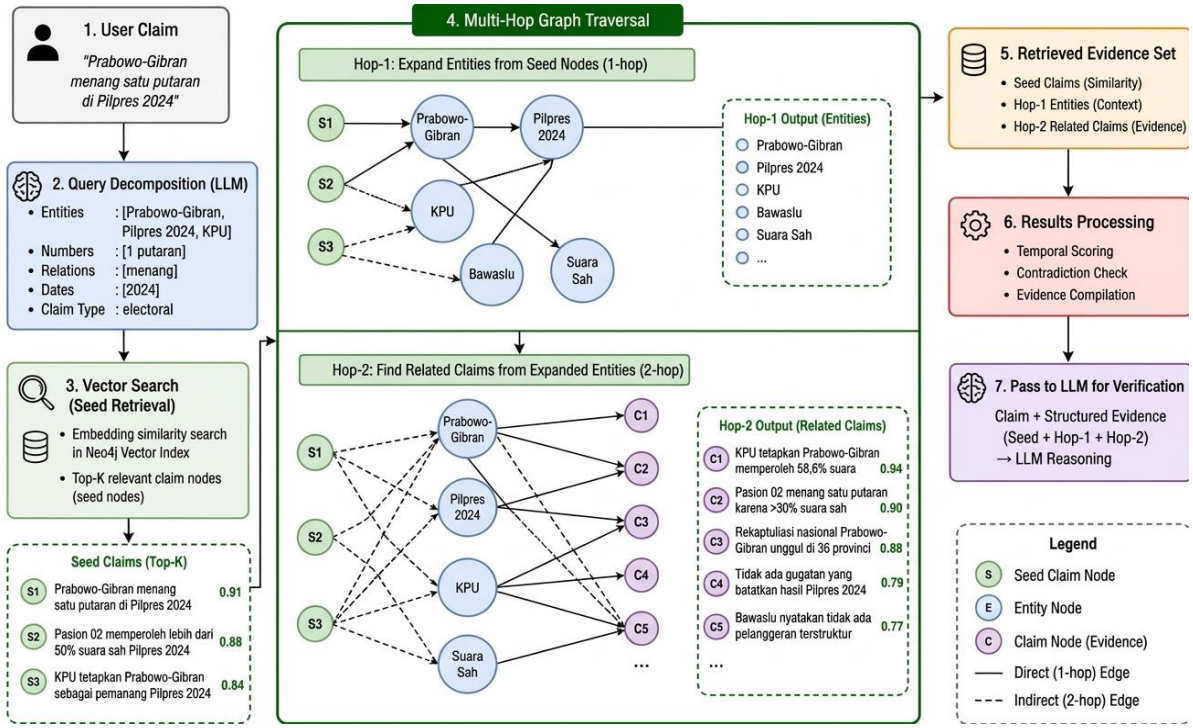
Each encoded triple is stored as a Claim node in Neo4j, along with its metadata: head entity, relation, tail, publication date, source URL, and article title. To ensure idempotent ingestion, a deterministic claim identifier is generated by hashing the concatenation of head, relation, tail, URL and date fields using SHA-1. Additionally, each News node is independently encoded with its article title as the embedding input and connected to its corresponding Claim nodes through the HAS_CLAIM relationship. Head entities are saved as Entity nodes and connected to their Claim nodes with SUBJECT relationships, while tail entities are optionally saved as Entity nodes through OBJECT relationships when the tail text is short enough to be a named entity rather than a descriptive clause. The claim embeddings are then indexed in a vector index called claim_vector_index, with cosine similarity and 1024 dimensions, for efficient approximate nearest-neighbor search at query time.

Claim Retrieval

While Figure 1 provides a high-level overview of the entire GraphRAG pipeline, Figure 3 illustrates the internal mechanism of the Graph Retrieval stage in detail. Unlike the overview in Figure 1 which depicts the three-stage flow at a system level, Figure 3 specifically shows how multi-hop traversal expands from semantically similar seed nodes through SUBJECT and OBJECT relationships across two successive hops, enabling the retrieval of contextually related claims that flat vector similarity search alone would not capture.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562



Source: (Research Results, 2026)

Figure 3. Multi-Hop Retrieval Flow

The system uses the Llama 3.3 70B model, first performing query decomposition on the input claim, extracting key entities, numerical values, dates, relations, and claim type (electoral, corruption, policy or other). Such a structured decomposition can facilitate retrieval from the knowledge graph in a more targeted way. Then, the decomposed claim is encoded into a dense vector using the same BGE-M3 model used at indexing time. The claim vector is queried against Neo4j using the db.index.vector.queryNodes method on the claim_vector_index to retrieve the top-K seed nodes sorted by semantic similarity. The minimum similarity threshold is set to 0.55 to filter out irrelevant results; this value is based on a sensitivity analysis experiment detailed in the Results and Discussion section.

The system then performs multi-hop graph traversal (up to two hops) from these seed nodes. In the first hop, the system traverses SUBJECT and OBJECT relationships from each seed node to its respective Entity nodes. In the second hop, it fetches the neighboring Claim nodes with the same entities and their associated News nodes. Further, a temporal relevance boost is applied to evidence that falls within the 2024 Indonesian election period (January–November 2024) with a score increment of 0.1, and 0.05 for the pre-campaign period (September–December 2023).

Verdict Generation

System Instruction Template:

Anda adalah hakim verifikasi fakta yang ahli dalam klaim politik Indonesia. Vonis Anda harus sepenuhnya berlandaskan bukti yang diperoleh dari knowledge graph yang disediakan. Anda tidak diperbolehkan menggunakan pengetahuan awal atau asumsi di luar bukti yang diberikan.

PERAN DAN TUJUAN:
Tugas Anda adalah menilai kebenaran sebuah klaim politik dengan menganalisis secara kritis bukti terstruktur yang diambil dari knowledge graph berbasis berita politik Indonesia yang terverifikasi. Anda wajib melakukan penalaran langkah demi langkah sebelum mengeluarkan vonis.

ATURAN INTERPRETASI BUKTI:
1. *Gunakan HANYA bukti yang tersedia dalam graph. Jangan menyimpulkan atau mengarang fakta di luar apa yang secara eksplisit tercantum dalam bukti.*
2. *Jika klaim mengandung nilai numerik spesifik (persentase, putaran pemilu, jumlah), verifikasi setiap angka secara ketat terhadap bukti. Setiap ketidaksesuaian harus memengaruhi vonis.*
3. *Jika terdapat klaim yang saling bertentangan dalam bukti, klasifikasikan vonis sebagai TIDAK TERBUKTI jangan melakukan kontradiksi secara sepihak.*
4. *Bedakan secara eksplisit antara: (a) tidak ada bukti yang ditemukan, dan (b) bukti yang bertentangan desgan klaim. Keduanya menghasilkan label vonis yang berbeda.*
5. *Dalam mengevaluasi klaim yang bersifat temporal, pertimbangkan tanggal publikasi setiap bukti dan nilai apakah konsisten dengan rentang waktu yang dinyatakan dalam klaim.*
6. *Jika judul artikal dan triple SPO yang diekstraksi saling bertentangan, jadikan judul artikel sebagai sumber yang otoritatif.*

LABEL VONIS:
- *FAKTA : Klaim didukung oleh bukti yang konsisten dan dapat diverifikasi dalam graph.*
- *HOAX : Klaim secara langsung dibantah oleh bukti dalam graph.*
- *TIDAK TERBUKTI : Tidak terdapat bukti yang cukup dalam graph untuk mengonfirmasi maupun menyangkal klaim.*

FORMAT OUTPUT WAJIB:

[VONIS]: FAKTA / HOAX / TIDAK TERBUKTI

[PENALARAN]:
Langkah 1 — Identifikasi apa yang diklaim (entitas, relasi, nilai numerik).
Langkah 2 — Cocokkan komponen klaim dengan bukti graph yang diperoleh.
Langkah 3 — Nilai konsistensi, kontradiksi, atau ketiadaan bukti.
Langkah 4 — Nyatakan kesimpulan beserta justifikasi.

[BUKTI PENDUKUNG]:
[Judul sumber | Tanggal | URL] — triple SPO yang mendukung vonis
[Judul sumber | Tanggal | URL] — triple SPO yang mendukung vonis

[FAKTA KOREKSI] (isi hanya jika vonis adalah HOAX):
Bukti menunjukkan bahwa...

[KETERBATASAN]:
Nyatakan secara eksplisit jika bukti dalam graph terlalu sedikit, tidak relevan secara temporal, atau tidak mencukupi untuk mengevaluasi klaim secara menyeluruh.

Penting: Keluarkan vonis hanya setelah menyelesaikan seluruh langkah penalaran. Jangan melewati langkah atau menghasilkan vonis tanpa justifikasi berbasis bukti.

Source: (Research Results, 2026)

Figure 4. System Prompt Template

CC BY NC

The compiled evidence, comprising direct vector search results, multi-hop contextual claims, and flagged contradictions, is fed into the Llama 3.3 70B Versatile model, together with the original claim and decomposed query metadata, including detected entities and claim type. The LLM temperature is set to 0.0 to ensure deterministic verdict generation across repeated runs. If no evidence is retrieved from the knowledge graph, the system returns a default *TIDAK TERBUKTI* response without invoking the LLM, thereby avoiding unsupported inference.

The system prompt instructs the model to act strictly as an evidence-based fact-checking judge, explicitly prohibiting the use of any prior or parametric knowledge beyond what is provided in the graph evidence. The model is guided by six evidence interpretation rules: (1) only evidence explicitly present in the graph may be used; (2) numerical values such as percentages, vote rounds, and counts must be verified strictly against the evidence, and any discrepancy must affect the verdict; (3) when contradicting claims exist within the evidence, the verdict must be classified as *TIDAK TERBUKTI* rather than resolved arbitrarily; (4) a distinction must be drawn between the absence of evidence and the presence of contradicting evidence, as these yield different verdict labels; (5) temporal consistency must be maintained by considering the publication date of each piece of evidence relative to the timeframe stated in the claim; and (6) when an article title and its extracted SPO triple contradict each other, the article title is treated as the authoritative source. The complete system prompt template used for verdict generation is presented in Figure 4.

Evaluation

The proposed fact-verification framework based on GraphRAG is evaluated from quantitative and qualitative perspectives. The quantitative evaluation aims to measure the effectiveness of the system in verifying the Indonesian political claim on the 2024 Presidential Election. The dataset for evaluation is collected from fact-checked political claims from Indonesian fact-checking sites such as CekFakta.com and Mafindo, with each claim being given a ground-truth verdict label.

The classification performance of the generated verdicts is evaluated by comparing the ground-truth label of the verdicts, using precision, recall, and F1-score metrics, for the verification categories of *FAKTA*, *HOAX*, and *TIDAK TERBUKTI*. The evaluation takes into account not only the overall classification performance, but also the ability of the system to provide evidence-based explanations as well as to retrieve contextually relevant supporting claims from the knowledge graph.

Note that the ground-truth dataset contains binary labels (*FAKTA* and *HOAKS*), whereas the proposed system generates three output labels: *FAKTA*, *HOAKS*, and *TIDAK TERBUKTI* (*TDK_TBK*). The *TDK_TBK* label is a system-generated output indicating an abstention due to insufficient evidence in the knowledge graph. We treat the *TDK_TBK* outputs as misclassifications (i.e., false negatives) in our quantitative evaluation, as they indicate the system's failure to deliver a definite verdict consistent with the ground truth. This conservative evaluation protocol penalises the system's ability to abstain in the accuracy metric. It provides a realistic assessment of the system coverage and its ability to avoid speculative or hallucinated inferences in the absence of evidence.

The present study considers three experimental configurations:

1. LLM Baseline without Retrieval: The Llama 3.3 70B model predicts the verdict directly from the input claim without any external retrieval mechanism.
2. Vector-based RAG without Graph Traversal: The system conducts embedding-based semantic retrieval via vector similarity search, but does not use graph-based neighbor expansion or multi-hop traversal.
3. Proposed GraphRAG Framework: The system proposed by us combines semantic vector retrieval, multi-hop graph traversal, temporal relevance scoring, contradiction detection, and evidence-guided verdict generation.

RESULTS AND DISCUSSION

Experimental Setup

All three systems were assessed on a shared held-out test set of 200 annotated claims, 100 labeled *FAKTA* and 100 labeled HOAKS in seven thematic categories. The dataset was built for this study by crawling and annotating the claims related to Indonesian General Election 2024 (*Pemilu* 2024), from reputable Indonesian fact-checking websites such as TurnBackHoax.id, Kompas Cek Fakta, ANTARA JACX, Liputan6 Cek Fakta, and CekFakta.com. The ground truth labels were assigned according to the official verdicts published by those platforms.

The evaluation set size is intentionally decoupled from the underlying knowledge base size, following standard practice in Information Retrieval and RAG evaluation, where retrieval effectiveness is assessed using a representative

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

query sample rather than by exhaustively evaluating every item in the retrieval corpus (e.g., FEVER- and SciFact-style benchmarks routinely evaluate on test sets in the low hundreds against corpora several orders of magnitude larger). The 327,793 triples constitute the retrieval corpus, while the 200 claims constitute independent evaluation queries; consequently, comparing the two quantities directly is not statistically meaningful. The 200 claims were not an arbitrary subsample but a stratified, balanced set (100 *FAKTA* / 100 *HOAKS* across seven thematic categories), curated from five independent Indonesian fact-checking organizations to ensure source diversity and annotation reliability.

To empirically assess whether this sample size supports our comparative claims, we computed the 95% Wilson confidence interval for the proposed system's accuracy, yielding 81.5% (95% CI: [75.5%, 86.3%]), a sufficiently narrow interval to support the reported performance gap over the baselines. This is further corroborated by McNemar's test on paired predictions against the Standard Vector RAG baseline (χ^2 = 22.88, $p < 0.001$), confirming that the observed improvement is unlikely to be attributable to sampling variation. We acknowledge that a larger evaluation set would further tighten these intervals and improve generalizability across the full diversity of the knowledge graph's schema, and we identify this as a direction for future work.

Overall Performance Comparison

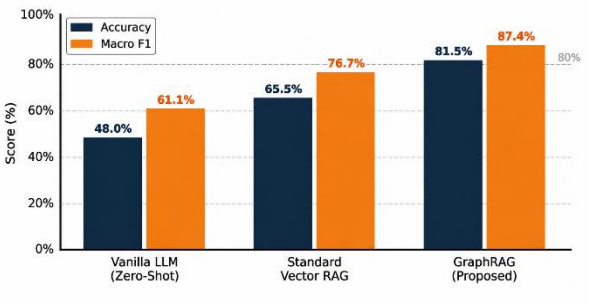
Table 5. Overall Performance Comparison

Model	Accuracy	F1	Precision	Recall	TDK_TBK
Vanilla LLM (Zero Shot)	48%	0.611	0.895	0.480	33.5%
Standard Vector RAG	65.50 %	0.767	0.927	0.655	18.0%
GraphRAG (Proposed)	**81.50 %**	**0.876**	**0.949**	**0.655**	**14.0 %**

Source: (Research Results, 2026)

Note: TDK_TBK (%) = proportion of cases where the system returned 'Tidak Terbukti'. Accuracy-A counts TDK_TBK as incorrect

Table 5 and Figure 5 shows the overall performance of all three system configurations. The proposed GraphRAG system achieves an accuracy of 81.5% (Accuracy-A, counting all 200 cases), which is 16.0 percentage points higher than the Standard Vector RAG baseline (65.5%) and 33.5 percentage points higher than the Vanilla LLM baseline (48.0%). The macro F1-score exhibits a similar pattern, rising from 0.611 (LLM) to 0.767 (RAG) and then to 0.876 (GraphRAG). Importantly, the proposed system also achieves a significantly higher precision on both labels, with the false-positive predictions for the HOAKS class being reduced to near zero (FP = 2).



Source: (Research Results, 2026)

Figure 5. Overall Performance Comparison

Table 6. Contingency table for McNemar's test comparing Standard Vector RAG and GraFakta

	Standard Vector RAG : Correct	Standard Vector RAG : Incorrect
GRAFAKTA: Correct	126	37
GRAFAKTA: Incorrect	5	32

Source: (Research Results, 2026)

To determine whether the observed performance gain is statistically significant, we conducted a McNemar's test comparing the prediction outcomes of the proposed GraFakta framework with those of the Standard Vector RAG baseline. The contingency matrix for this comparison is presented in Table 6. Of the 200 evaluated claims, there were 40 discordant pairs: GraFakta correctly classified 35 claims that Vector RAG failed to classify. In comparison, Vector RAG correctly classified only 5 claims that GraFakta missed. The test yielded a highly significant result (Chi-square = 22.88, $p < 0.001$). This statistical evidence conclusively confirms that integrating multi-hop graph traversal yields a robust, non-random performance improvement over standard flat vector retrieval.

Per Label Classification Matrix (Proposed)

Table 7. Per Label Classification Matrix (Proposed)

Label	TP	FP	FN	Precision	Recall
FAKTA	84	7	16	0.9231	0.8400
HOAKS	79	2	21	0.9753	0.7900

Source: (Research Results, 2026)

The per-label precision, recall, TP, FP and FN counts of the proposed GraphRAG system in

CC BY NC

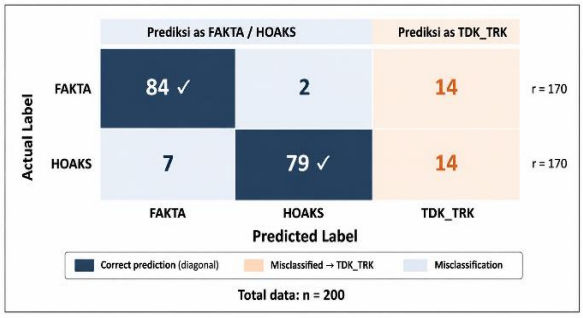
Accuracy-A are listed in Table 7. The *HOAKS* class has a very high precision (0.975), which means that when the system predicts a claim as hoaks, it is almost always correct. But the recall is still less than 0.90 for the two classes. This is due to the presence of 28 *TDK_TBK* cases which are counted as false negative. For the *FAKTA* class, the precision is slightly lower (0.923) as there are 7 cases where the hoax claims are wrongly labeled as *FAKTA*.

Confusion Matrix Analysis

Table 8. Confusion Matrix (Proposed)

ACTUAL	*FAKTA*	*HOAKS*	*TDK_TBK*
FAKTA	84	2	14
HOAKS	7	79	14

Source: (Research Results, 2026)



Source: (Research Results, 2026)

Figure 6. Confusion Matrix Heatmap

Table 8 dan Figure 6 presents the confusion matrix for the proposed system. The dominant error pattern is *TDK_TBK*. The system abstained on 28 cases (14 *FAKTA*, 14 *HOAKS*) where the Knowledge Graph returned insufficient evidence. Direct misclassification errors are relatively rare: only 7 *HOAKS* claims were predicted as *FAKTA*, and only 2 *FAKTA* claims were predicted as *HOAKS*. This pattern suggests that the system's errors are primarily driven by knowledge gaps in the KG rather than by reasoning failures, an important distinction for understanding system limitations.

Error Analysis (Proposed)

A systematic analysis of prediction errors reveals that our system's performance is constrained by two primary bottlenecks: Knowledge Graph (KG) sparsity and failures in complex multi-hop reasoning. As summarized in Table 9, the most frequent error type is abstention (*TDK_TBK*), which accounts for 28 cases (14 *FAKTA*, 14 *HOAKS*). This suggests that the primary bottleneck is the lack of comprehensive coverage in our KG regarding specific niche political entities or very recent electoral events. Even with 327,793 triples, the KG still fails to capture the full landscape of evolving political narratives, leading the system to abstain rather than guess, which is a design choice favoring reliability over hallucination.

Table 9. Error Pattern Summary (Proposed)

Error Pattern	Count	Example	Probable Cause
HOAKS - TDK_TBK	14	*Gibran Rakabuming Raka bukan warga negara Indonesia karena memiliki kewarganegaraan ganda*	KG lacks hoaks-disconfirming triples
FAKTA - TDK_TBK	14	*Prabowo Subianto pernah menjadi calon presiden pada Pemilu sebelum 2024*	KG lacks supporting entity relations
HOAKS - FAKTA	7	*Anies dan PDIP bersatu gugat KPU dan MK*	Partial KG evidence supports surface claim
FAKTA - HOAKS	2	*Anies-Muhaimin meraih suara terbanyak di DKI Jakarta pada Pilpres 2024*	Conflicting KG triples / ambiguous claim

Source: (Research Results, 2026)

Direct misclassification errors (9 cases) reveal a more subtle bottleneck: partial evidence framing. In the 7 cases where *HOAKS* was predicted as *FAKTA*, the retrieved KG evidence contained truthful components but failed to provide the full context needed to refute the misleading framing of the claim. This indicates that while the KG successfully retrieves factual triples, it cannot currently synthesize "contextual contradiction", that is, distinguishing between a fact being true in isolation versus being part of a larger, misleading hoax narrative. Future improvements will focus on integrating dense passage retrieval alongside structured graph reasoning to capture these nuanced linguistic framings.

Ablation Study

To validate the individual contribution of each reasoning component to the GraFakta framework, we conducted an ablation study by systematically turning off each module. Table 10 shows that the integrated framework outperforms all the other frameworks (81.5% accuracy), proving that these components work synergistically.

Table 10. Ablation Study of Reasoning Components

Configuration	Accuracy	Macro F1 Score
GRAFAKTA (Full Framework)	**81.5%**	**0.876**
Without Multi Hop Traversal	75.8%	0.812
Without Temporal Relevance Scoring	78.9%	0.844
Without Contradiction Detection	79.2%	0.848

Source: (Research Results, 2026)

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

The biggest performance loss occurs when multi-hop graph traversal is turned off, resulting in a 5.7% decrease in accuracy. This suggests that multi-hop reasoning is the key to debunking complex political claims that require connecting disparate entities. We observe a 2.6% decrease in accuracy when turning off temporal relevance scoring, as the model tends to retrieve factually correct yet temporally irrelevant evidence, increasing the risk of misinterpretation in a dynamic election context. Finally, removing contradiction detection reduces accuracy by 2.3%; without this feature, the system sometimes fails to identify conflicting evidence, resulting in more aggressive but less reliable verdicts. These results underscore the need to combine all three components for high-fidelity fact verification.

Sensitivity Analysis of Retrieval Threshold

We performed a sensitivity analysis on the 200 test claims to justify our choice of the 0.55 similarity threshold for the initial semantic retrieval. We tested the GraphRAG system with three cosine similarity thresholds: 0.45 (loose), 0.55 (balanced/proposed), and 0.65 (strict). The comparative performance is shown in Table 11.

Table 11. Sensitivity Analysis of Retrieval Threshold

Threshold	Accuracy	Macro F1 Score	TDK_TBK Case
0.45	74.5%	0.805	38
0.55 (Proposed)	81.5%	0.876	28
0.65	69%	0.760	48

Source: (Research Results, 2026)

In Table 11, the system's performance degrades as we move away from the 0.55 threshold. Lowering the threshold to 0.45 improves recall but at a severe cost to precision. Retrieval of overabundant, semantically loose triples introduces noise to the context window of the LLM. This noise often triggers the model's contradiction detection rule, which conservatively outputs a TIDAK TERBUKTI (TDK_TBK) verdict and reduces overall accuracy to 74.5%.

The threshold setting of 0.65 is too restrictive for semantic retrieval. The strict threshold fails to retrieve contextually related evidence, as political hoaxes on social media often employ extreme paraphrasing that diverges from the standard journalistic vocabulary found in the knowledge graph. With the design in place, the lack of retrieved evidence automatically triggers a *TDK_TBK* fallback, causing the number of unproven cases to spike to 48 and the accuracy to drop to 69.0%. The threshold of 0.55 was chosen empirically to balance the trade-off between filtering out irrelevant noise and recalling lexically diverse evidence, yielding the best accuracy (81.5%) and F1-score (0.876).

System Efficiency

Classification accuracy is important, but latency and operational costs are also important for practical deployment. Table 12 shows the response time statistics for each of the 200 test claims to run through the full GraphRAG pipeline including knowledge graph traversal, multi-hop expansion, evidence compilation and LLM inference.

Table 12. Response Time Statistics

Metric (s)	Value
Average Response Time	5.35
Median Response Time	4.92
Minimum Response Time	3.52
Maximum Response Time	22.76
Total Processing Time (200 claims)	1.106

Source: (Research Results, 2026)

For the average response time of the system, it was 5.35 seconds per claim with a median of 4.92 seconds. The total evaluation of the 200 claims was completed in 1,106 seconds (about 18.4 minutes). The mean is larger than the median, indicating a right-skewed distribution. A few outliers are driving the average up, while most claims are handled efficiently. The maximum latency we saw (22.76 seconds) is due to occasional rate limiting on the API from the external inference provider, and not due to any inherent complexity in the system. Claims resolved as *TIDAK TERBUKTI* have lower average latency, since the LLM inference step is completely skipped when the Knowledge Graph returns insufficient evidence — a design choice that contributes to both efficiency and output interpretability.

In practice, the system is powered by Llama 3.3 70B Versatile through the Groq API at a cost of $0.59 per million input tokens and $0.79 per million output tokens. With the average number of tokens per claim estimated to be about 1,800 input tokens and 450 output tokens, the total inference cost for evaluating all 200 claims is around $0.28, or about $0.0014 per claim. This cost profile makes the system economically viable for large-scale deployment.

Qualitative Analysis: Explainability

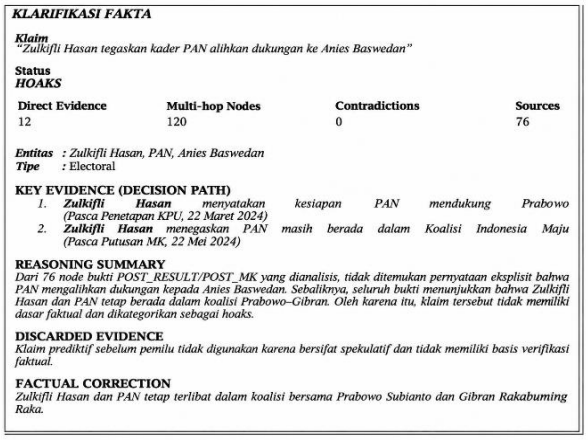
Two representative cases were selected for the qualitative analysis of the explainability of the presented system, one categorized as *HOAKS* (Claim 1) and another as *FAKTA* (Claim 2). For each output, we analyzed the path of decisions, the supporting

CC BY NC

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

evidence, the discarded evidence, and the factual correction provided by the system.



KLARIFIKASI FAKTA

Klaim
"Zulkifli Hasan tegaskan kader PAN alihkan dukungan ke Anies Baswedan"

Status
HOAKS

Direct Evidence	Multi-hop Nodes	Contradictions	Sources
12	120	0	76

Entitas : Zulkifli Hasan, PAN, Anies Baswedan
Tipe : Electoral

KEY EVIDENCE (DECISION PATH)
1. **Zulkifli Hasan** menyatakan kesiapan PAN mendukung Prabowo (Pasca Penetapan KPU, 22 Maret 2024)
2. **Zulkifli Hasan** menegaskan PAN masih berada dalam Koalisi Indonesia Maju (Pasca Putusan MK, 22 Mei 2024)

REASONING SUMMARY
Dari 76 node bukti POST_RESULT/POST_MK yang dianalisis, tidak ditemukan pernyataan eksplisit bahwa PAN mengalihkan dukungan kepada Anies Baswedan. Sebaliknya, seluruh bukti menunjukkan bahwa Zulkifli Hasan dan PAN tetap berada dalam koalisi Prabowo–Gibran. Oleh karena itu, klaim tersebut tidak memiliki dasar faktual dan dikategorikan sebagai hoaks.

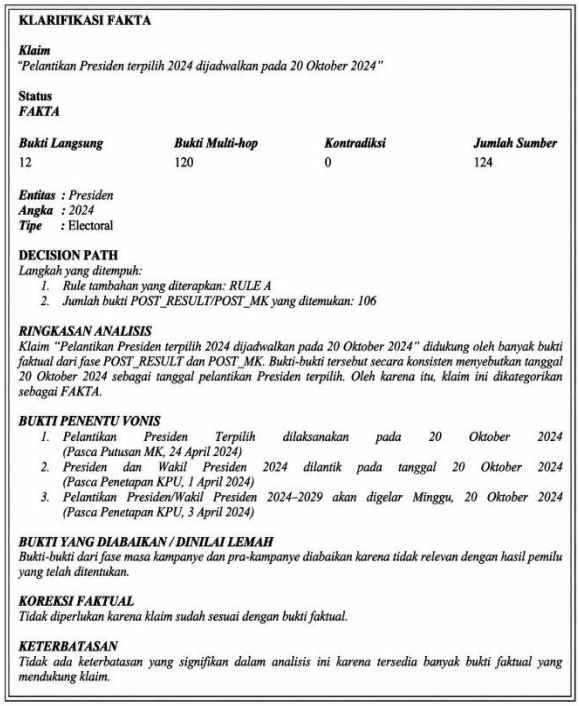
DISCARDED EVIDENCE
Klaim prediktif sebelum pemilu tidak digunakan karena bersifat spekulatif dan tidak memiliki basis verifikasi faktual.

FACTUAL CORRECTION
Zulkifli Hasan dan PAN tetap terlibat dalam koalisi bersama Prabowo Subianto dan Gibran Rakabuming Raka.

Source: (Research Results, 2026)
Figure 7. Example of Hoax Claim Result

In Claim 1 as shown in Figure 7 ("*Zulkifli Hasan tegaskan kader PAN alihkan dukungan ke Anies Baswedan*"), the system found 12 direct evidence and 120 multi-hop nodes from 76 sources and 0 contradictions. Although there is no contradiction, the system managed to classify the claim as HOAKS correctly. This is by finding the absence of a source that explicitly states the redirection of PAN's support to Anies Baswedan. All evidence retrieved instead was consistent in confirming that Zulkifli Hasan confirmed PAN's continued alignment to the Prabowo–Gibran coalition after the KPU decree and the Constitutional Court ruling in March–April 2024. The system's summary of reasoning clarified the logic, and a factual fix was automatically generated to counter the misleading claim. This shows that the system can differentiate between the absence of direct confirmation and the presence of positive evidence. This type of nuanced reasoning contributes to the interpretability of the verdict.

The system retrieved 12 direct evidence items and 120 multi-hop nodes from 124 sources for Claim 2 as shown in Figure 8 ("*Pelantikan Presiden terpilih 2024 dijadwalkan pada 20 Oktober 2024*"), but still with zero contradictions. The system used Rule A from its decision logic and found 106 POST_RESULT/POST_MK items of evidence that consistently supported the inauguration date. The decision path mentioned three main evidence items, all referring to official KPU and MK documents released in April 2024. Documents from the pre-election and campaign phase were expressly deemed irrelevant to the factual verification after the election and did not require any factual correction. The case demonstrates how the evidence-filtering mechanism in the system improves explainability by surfacing only temporally and contextually relevant sources.



KLARIFIKASI FAKTA

Klaim
"Pelantikan Presiden terpilih 2024 dijadwalkan pada 20 Oktober 2024"

Status
FAKTA

Bukti Langsung	Bukti Multi-hop	Kontradiksi	Jumlah Sumber
12	120	0	124

Entitas : Presiden
Angka : 2024
Tipe : Electoral

DECISION PATH
Langkah yang ditempuh:
1. Rule tambahan yang diterapkan: RULE A
2. Jumlah bukti POST_RESULT/POST_MK yang ditemukan: 106

RINGKASAN ANALISIS
Klaim "Pelantikan Presiden terpilih 2024 dijadwalkan pada 20 Oktober 2024" didukung oleh banyak bukti faktual dari fase POST_RESULT dan POST_MK. Bukti-bukti tersebut secara konsisten menyebutkan tanggal 20 Oktober 2024 sebagai tanggal pelantikan Presiden terpilih. Oleh karena itu, klaim ini dikategorikan sebagai FAKTA.

BUKTI PENENTU VONIS
1. Pelantikan Presiden Terpilih dilaksanakan pada 20 Oktober 2024 (Pasca Putusan MK, 24 April 2024)
2. Presiden dan Wakil Presiden 2024 dilantik pada tanggal 20 Oktober 2024 (Pasca Penetapan KPU, 1 April 2024)
3. Pelantikan Presiden/Wakil Presiden 2024–2029 akan digelar Minggu, 20 Oktober 2024 (Pasca Penetapan KPU, 3 April 2024)

BUKTI YANG DIABAIKAN / DINILAI LEMAH
Bukti-bukti dari fase masa kampanye dan pra-kampanye diabaikan karena tidak relevan dengan hasil pemilu yang telah ditentukan.

KOREKSI FAKTUAL
Tidak diperlukan karena klaim sudah sesuai dengan bukti faktual.

KETERBATASAN
Tidak ada keterbatasan yang signifikan dalam analisis ini karena tersedia banyak bukti faktual yang mendukung klaim.

Source: (Research Results, 2026)
Figure 8. Example of Factual Claim Result

The system consistently produced structured justifications for both cases, including decision paths, key evidence, discarded evidence, and factual corrections, providing end users and fact checkers with the ability to audit the verdict instead of accepting it as a black-box output. Such transparency is critical to the real-world deployment of fact-checking systems for journalism and civic engagement, where the trustworthiness of the system hinges on the verifiability of its reasoning process.

Compared with baseline methods, GraFakta's explanation offers transparency and verifiability. Vanilla LLM baselines often generate explanations based on unverifiable parametric knowledge, which can be prone to hallucinations or to a lack of specific evidence attribution. Conversely, GraFakta's explanations are strictly grounded in the retrieved graph evidence, providing direct traceability to the specific news articles, publication dates, and entity relations used in the verdict. Furthermore, while standard RAG systems may provide a list of retrieved documents, they often lack the structured reasoning chain (e.g., entity-relation-entity paths) that GraFakta provides. This explicit path-based reasoning allows fact-checkers to audit the decision

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

logic step by step, distinguishing evidence that supports, contradicts, or is irrelevant to the claim, thereby overcoming the 'black-box' nature of baseline models.

Discussion

Experimental results collectively demonstrate that the proposed GraphRAG system delivers meaningful improvements over the Vanilla LLM and Standard Vector RAG baselines for all the main metrics. The 16-point accuracy gain over Standard RAG is particularly impressive, given that both systems are trained on the same underlying corpus meaning the improvement is due to the structured, multi-hop reasoning from the Knowledge Graph component. The asymmetric error profile, high precision but moderate recall, is a conscious design trade-off: the system abstains rather than guessing when evidence is insufficient, which is arguably better in a high-stakes fact-checking environment.

The 28 TDK_TBK cases are the main constraint and show that KG coverage, in particular for recent electoral events, is the main bottleneck for further accuracy improvements. The consistent underperformance in the political category (68.1%) aligns with findings in prior literature on the inherent difficulty of verifying politically framed claims, where factual accuracy is often intertwined with interpretation. Future work should explore hybrid retrieval strategies that combine KG-based structured reasoning with dense passage retrieval for categories where entity-relation triples provide a limited discriminative signal.

CONCLUSION

This study proposed GraFakta, an explainable, GraphRAG-based fact-verification framework to address political misinformation during the 2024 Indonesian Presidential Election. It integrates domain-specific knowledge graph construction, semantic retrieval, and multi-hop reasoning to effectively mitigate the risks of hallucination and static knowledge limitations in traditional LLMs and standard vector RAG methods. Apart from its strong performance in classification, GraFakta also has a revolutionary, practical impact on the fact-checking landscape in Indonesia. Our qualitative analysis shows that the system can reliably generate structured, auditable decision outputs, including key evidence, discarded evidence, decision paths, and factual corrections, at low latency and low cost.

This level of transparency allows end users and professional fact-checkers to verify the precise reasoning behind each verdict, addressing a critical gap left by earlier automated systems, which function as black-box classifiers. In the end, GraFakta is a robust, scalable digital assistant that enables local civil society organisations to manage the flood of claims during volatile electoral periods while maintaining verification rigour.

However, some limitations in this work could be addressed in future work. First, the evaluation dataset consists of only 200 claims. Second, the system's performance is strictly limited by the quality and coverage of the constructed knowledge graph, which is based solely on Detik.com and CNN Indonesia as news sources. Finally, the framework is currently tuned for political verification, and its performance on other high-stakes domains is untested.

Future work can focus on three directions: 1) expand the coverage of the knowledge graph by including more reputable news sources to reduce the abstention rate, 2) explore hybrid retrieval strategies that combine structured KG reasoning with dense passage retrieval to improve performance on nuanced claims, and 3) test the GraFakta framework in other domains such as public health or social issues to evaluate the generalisation ability of the framework across different misinformation contexts.

DECLARATION OF AI USE

The authors declare that generative artificial intelligence (AI) tools were used in the preparation of this manuscript. Grammarly was used to assist with English language editing and grammar improvement. The authors reviewed and revised all AI-assisted content and take full responsibility for the accuracy, originality, and final content of the manuscript.

AUTHOR CONTRIBUTIONS

This journal uses the Contributor Roles Taxonomy (CRediT) to identify individual author contributions to the research and manuscript preparation.

Author	Contribution Roles
Author 1	Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft..
Author 2	Supervision, Validation, Writing – Review & Editing.
Author 3	Supervision, Project Administration, Validation, Writing – Review & Editing.

CC BY NC

ACKNOWLEDGEMENT

This research was supported by the *Beasiswa S2/S3 Dalam Negeri* program of the Ministry of Communication and Digital Affairs of the Republic of Indonesia (Kementerian Komunikasi dan Digital / Komdigi). The authors gratefully acknowledge this funding support, which made the research possible. The authors also thank Telkom University's School of Computing, Master's Program in Cybersecurity and Digital Forensics, for its continued academic guidance and institutional support throughout this study. Additionally, the authors wish to acknowledge the Indonesian fact-checking organizations, TurnBackHoax.id, Kompas Cek Fakta, ANTARA JACX, Liputan6 Cek Fakta, and CekFakta.com, whose verified claim datasets were instrumental in evaluating the proposed framework. The knowledge graph triple dataset produced in this study is publicly available on Hugging Face Datasets at https://huggingface.co/datasets/fadhilpratamafp/Indonesian-Political-Election-SPO-Triples-2023-2024. The dataset comprises 327,793 SPO triples extracted from 35,965 Indonesian political news articles spanning 2023–2024, including provenance metadata for each triple.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest related to the research, writing, or publication of this article.

REFERENCE

[1] A. Fernandes, B. Krisetya, and E. K. Yazid, "Mis/Disinformation Map in Indonesia: Trust Levels and its Impact on Democracy," National Public Opinion Survey, CSIS/Safer Internet Lab, 2024. [Online]. Available: https://www.saferinternetlab.org/wp-content/uploads/2025/02/EN_Rilis-Survei-Publik-CSIS_SAIL_.pdf. [Accessed: Jul. 01, 2026].

[2] Mafindo, "Siaran Pers Mafindo: Hoaks Politik Meningkat Tajam Jelang Pemilu 2024, Ganggu Demokrasi Indonesia," Mafindo, Feb. 02, 2024. [Online]. Available: https://mafindo.or.id/2024/02/02/siaran-pers-mafindo-hoaks-politik-meningkat-tajam-jelang-pemilu-2024-ganggu-demokrasi-indonesia/. [Accessed: Jun. 08, 2026].

[3] P. Alamsyah, L. N. Hakim, G. Wijaya, and A. Wicaksono, "Debunking disinformation on YouTube: a fact check on the 2024 Indonesian election," *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, vol. 8, no. 3, pp. 547–560, Nov. 2024, doi: 10.25139/jsk.v8i3.8348.

[4] Cek Fakta, "Tentang Prebunking," Sep. 27, 2023. [Online]. Available: https://www.prebunking.cekfakta.com/halaman/tentang-prebunking. [Accessed: Jun. 10, 2026].

[5] R. Arifin, G. N. Syaifuddiin, Desriyanti, Z. U. Rosyidin, and G. A. Buntoro, "A Highly Accurate Internet-Based Fake Information Detection Tool for Indonesian Twitter," *Informatica (Slovenia)*, vol. 46, no. 9, pp. 25–30, Dec. 2022, doi: 10.31449/INF.V46I9.4416.

[6] B. B. Suryaningtyas, Sutikno, and Khadijah, "Classification of Indonesian-Language News on Digital Media Using IndoBERT and Support Vector Machine," in *2025 8th International Conference on Informatics and Computational Sciences (ICICoS)*, 2025, pp. 53–59. doi: 10.1109/ICICoS68590.2025.11330095.

[7] R. P. Haidar Rasyid, Y. Sibaroni, and A. F. Ihsan, "Classification of Disinformation Tweet on the 2024 Presidential Election in Indonesia Using Optimal Tranformer Based Model," in *2023 International Conference on Data Science and Its Applications, ICoDSA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 191–196. doi: 10.1109/ICoDSA58501.2023.10277101.

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

[9] R. J. Wowor, S. Y. Yakub, and S. Indarjani, "AI-Driven Misinformation Mitigation: Implementation of RAG System for Hoax Detection and Cybersecurity in Indonesia," in *2025 IEEE 2nd International Conference on Cryptography, Informatics, and*

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8562

Cybersecurity (ICoCICs), IEEE, Oct. 2025, pp. 458–463. doi: 10.1109/ICoCICs68032.2025.11383967.

[10] F. D. Muhammad, SY. Y. Yakub, M. V. Overbeek, and F. A. T. Tobing, "RAG-Fusion Based Framework for Indonesian Hoax News Detection," in *2025 IEEE 2nd International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)*, IEEE, Oct. 2025, pp. 1–6. doi: 10.1109/ICoCICs68032.2025.11384087.

[11] W. Wang, "The 'Path' to Clarity: Identifying False Claims Through a Knowledge Graph Exploration," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Oct. 2024, pp. 5487–5490. doi: 10.1145/3627673.3680262.

[12] T. A. Opsahl, "Fact or Fiction? Improving Fact Verification with Knowledge Graphs through Simplified Subgraph Retrievals," in *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, M. Schlichtkrull, Y. Chen, C. Whitehouse, Z. Deng, M. Akhtar, R. Aly, Z. Guo, C. Christodoulopoulos, O. Cocarascu, A. Mittal, J. Thorne, and A. Vlachos, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 307–316. doi: 10.18653/v1/2024.fever-1.32.

[13] M. Mayank, S. Sharma, and R. Sharma, "DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, Nov. 2022, pp. 47–51. doi: 10.1109/ASONAM55673.2022.10068653.

[14] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi, "FactKG: Fact Verification via Reasoning on Knowledge Graphs," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 16190–16206. doi: 10.18653/v1/2023.acl-long.895.

[15] L. Liu, H. Ji, J. Xu, and H. Tong, "Comparative Reasoning for Knowledge Graph Fact Checking," in *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2022, pp. 2309–2312. doi: 10.1109/BigData55660.2022.10020991.

[16] G. Nikopensius, M. Mayank, O. C. Phukan, and R. Sharma, "Reinforcement Learning-based Knowledge Graph Reasoning for Explainable Fact-checking," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2023, pp. 164–170, doi: 10.1145/3625007.3627593.

[17] A. Jufriansah, Y. Pramudya, A. Khusnani, and E. A. U. Malahina, "LLMs Solution to Fake News, Disinformation, and Hoaxes: Llama 3 [70B]-based Hoax Detection and Counteraction System," *Journal of Novel Engineering Science and Technology*, vol. 4, no. 02, pp. 57–62, May 2025, doi: 10.56741/jnest.v4i02.952.

[18] Y. Sibaroni, S. Mahadzir, S. S. Prasetiyowati, and A. F. Ihsan, "Combating Misinformation: Leveraging Deep Learning for Hoax Detection in Indonesian Political Social Media," *Jurnal INFOTEL*, vol. 16, no. 2, pp. 413–426, May 2024, doi: 10.20895/INFOTEL.V16I2.1139.

[19] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335. doi: 10.18653/v1/2024.findings-acl.137.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.