Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 4 (2025) 729 - 736 e-ISSN: 2580-0760

DiG-MFV: Dual-integrated Graph for Multilingual Fact Verification

Nova Agustina[1*], Kusrini[2], Ema Utami[3], Tonny Hidayat[4]
[1,2,3,4]Department of Informatics Doctorate, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

[1]nova@students.amikom.ac.id, [2]kusrini@amikom.ac.id, [3]ema.u@amikom.ac.id, [4]tonny@amikom.ac.id

Abstract

The proliferation of misinformation in political domains, especially across multilingual platforms, presents a major challenge to maintaining public information integrity. Existing models often fail to effectively verify claims when the evidence spans multiple languages and lacks a structured format. To address this issue, this study proposes a novel architecture called Dual-integrated Graph for Multilingual Fact Verification (DiG-MFV), which combines semantic representations from multilingual language models (i.e., mBERT, XLM-R, and LaBSE) with two graph-based components: an evidence graph and a semantic fusion graph. These components are processed through a dual-path architecture that integrates the outputs from a text encoder and a graph encoder, enabling deeper semantic alignment and cross-evidence reasoning. The PolitiFact dataset was used as the source of claims and evidence. The model was evaluated by using a data split of 70% for training, 20% for validation, and 10% for testing. The training process employed the AdamW optimizer, cross-entropy loss, and regularization techniques, including dropout and early stopping based on the F1-score. The evaluation results show that DiG-MFV with LaBSE achieved an accuracy of 85.80% and an F1-score of 85.70%, outperforming the mBERT and XLM-R variants, and proved to be more effective than the DGMFP baseline model (76.1% accuracy). The model also demonstrated stable convergence during training, indicating its robustness in cross-lingual political fact verification tasks. These findings encourage further exploration in graph-based multilingual fact verification systems.

Keywords: fact verification; graph fusion; LaBSE; multilingual model; mBERT; political claim; XLM-R

How to Cite: [N. Agustina, Kusrini, E. Utami, and T. Hidayat, "DiG-MFV: Dual-integrated Graph for Multilingual Fact Verification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 4, pp. 729 - 736, Jul. 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i4.6695

Received: May 30, 2025
Accepted: July 14,2025
Available Online: July 27, 2025

1. Introduction

Fact verification has become a critical component in maintaining the integrity of public information in the current era of disinformation, particularly in the political domain. The increasing spread of false claims and hoaxes has driven the development of automated models based on Natural Language Processing (NLP) [1]-[4] to perform verification tasks quickly and accurately. Various approaches have been developed, ranging from Natural Language Inference (NLI) models [5] to pre-trained language models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and A Robustly Optimized BERT Pretraining Approach (RoBERTa) [6]. In recent years, efforts to combine semantic context and relational structures between evidences have led to advances through graph-based approaches [5]. However, existing systems face two critical limitations, i.e., they struggle with unstructured political claims in natural language format, and lack effective integration of multilingual evidence despite available cross-lingual models.

The combination of multiple models has been proven to improve performance in classification tasks [7]-[9]. Several studies have combined models, whether among transformer-based models [10], [11], among graph-based models [12], [13], or a hybrid between transformer and graph-based architectures [14]. This integration is based on the complementarity of representations: transformer-based models excel in capturing semantic information through pre-trained embeddings [15], while graph models are effective in representing relational structures between nodes [16]. these two types of representations are combined, the model can simultaneously understand sentence meaning (semantic matching) and inter-evidence relationships (relational reasoning), as demonstrated in previous studies, which show that this approach enables deeper and more accurate inference [14], particularly in

fact verification structures that require cross-lingual contextual understanding.

One of the models that integrates graph structures for fact verification tasks is the Double Graph Attention Network Reasoning Method (DGMFP) [13], which has demonstrated promising results in table-based fact verification. DGMFP incorporates a monolingual transformer, i.e., BERT, to capture the semantic meaning of claims and evidence, and employs a double graph mechanism to represent and reason over relationships among evidence. However, this approach still has several limitations. First, DGMFP is specifically designed for semi-structured data, making it less flexible when applied to facts presented in natural sentences without explicit tabular structure. Second, the use of logical forms as symbolic evidence heavily relies on the performance of a semantic parser, which in many cases is weakly supervised and prone to producing noisy or inaccurate programs. Third, although DGMFP integrates two types of graphs to enhance reasoning, the model has not yet optimized the contextual representation power of multilingual language models, which have proven effective in understanding cross-lingual claims. However, the model's ability to generalize and adapt in multilingual political fact verification remains limited, and its interpretability has not been fully integrated with the semantics of natural text. Currently, there is no double graph approach that explicitly integrates the power of multilingual contextual embeddings with semantic relationship-based evidence filtering and fusion mechanism in the political domain. Furthermore, the conclusion of these limitations of DGMFP has several weaknesses, including (1) the lack of effective multilingual evidence integration despite the availability of cross-lingual models, and (2) failure to test the combination of structural and semantic reasoning within a unified framework.

Previous studies on fact verification have primarily focused on monolingual models [17]-[20], and the exploration of combining two semantic graphs within the context of naturalistic political news text has not yet been conducted. In addition, no existing study has systematically evaluated and compared the contributions of various multilingual models such as Multilingual Bidirectional Encoder Representations from Transformers (mBERT)[21], Language agnostic BERT Sentence Embedding (LaBSE) [22], and Cross lingual Language Model RoBERTa (XLM-R) [23] within a double graph framework to improve the classification accuracy of political claims.

This study aims to develop and evaluate a novel architecture called Dual Integrated Graph for Multilingual Fact Verification (DiG-MFV). Unlike previous models that rely on monolingual encoders or symbolic and tabular structures, DiG-MFV integrates multilingual contextual embeddings from mBERT, LaBSE, and XLM-R to handle unstructured political claims across multiple languages. The proposed model introduces a dual-graph reasoning framework that combines two complementary components: (1) an evidence graph, which captures inter-evidence relationships based on co-occurrence and semantic similarity, and (2) a semantic fusion graph, which integrates claim-evidence pairs into a unified contextual space to support deeper semantic reasoning. These two graph-based representations are combined through a joint reasoning module that enables the model to perform both relational and semantic inference. This unified structure allows DiG-MFV to address several limitations found in previous studies, such as the lack of multilingual capability, inflexibility in processing natural language input, and the separation between semantic and structural reasoning components. The classification output generated by the model consists of three possible labels: True (if the statement is supported by the evidence), False (if the statement is contradicted by the evidence), or Not Enough Info (if the evidence is insufficient to support or refute the statement). This integrated approach represents the main contribution of this research by offering a flexible, multilingual, and interpretable fact verification framework suitable for real-world political discourse.

2. Methods

This section describes the method used in the development and evaluation of the Dual integrated Graph for Multilingual Fact Verification (DiG-MFV) model. The model combines multilingual language representations with a dual graph structure to perform political fact verification based on multiple evidences. The algorithmic approach developed from DGMFP, which originally used traditional double graphs i.e., GAT and KGAT for tabular data analysis. The difference with the research developed by the authors lies in replacing KGAT in DGMFP with cross-lingual evidence graphs based on multilingual embeddings, and enhancing the traditional GAT with a semantic fusion graph that processes in parallel, specifically the structural relations between claims and evidence.

2.1 DIG-MFV

The DiG-MFV (Dual integrated Graph for Multilingual Fact Verification) model is designed to perform fact classification of political claims using a graph-based representation approach combined with multilingual language models. The DiG-MFV system begins by receiving an input consisting of a single claim in sentence form and several candidate evidences support or refute the claim. Among all candidate evidences, the model selects the most relevant ones based on their semantic similarity to the claim. For example, evidence that states there is no increase in crime may be selected as the top-1 evidence. The selected evidences range from top-1 to top-6 based on prior findings [13], which indicate that the model achieves optimal sensitivity when using six evidences. After the claim and top-k

evidences are processed in the Instruction Style Concatenation block, the workflow splits into two parallel paths, i.e., the text encoder and the graph encoder. In the text encoder, the claim and selected evidences are structured in an instruction-like format similar to a question-answering prompt, in order to help the encoder capture the logical context between sentences. The workflow of this method is illustrated in Figure 1.

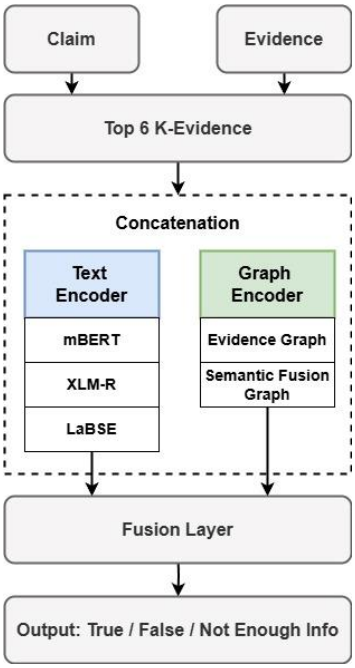

Figure 1. Architecture of the Proposed DiG-MFV Model

The output used is the final representation (h_{text}) of the [CLS] token, which is a special token placed at the beginning of each input in models such as mBERT, XLM-R, or LaBSE. The [CLS] token does not represent any specific word in the sentence, but instead serves as an aggregated representation of the entire input.

In the Graph Encoder, the embedding representations of the evidences obtained from the text encoder are averaged (average pooled), and then processed through a two-layer neural network module to produce a graph representation (z_{graph}), when the Evidence Graph (k) connects the embeddings (h) of the selected evidences, while the Semantic Fusion Graph (Enc_{graph}) is used to capture the semantic relationships among the evidence elements. The computation of the Graph Encoder is illustrated in Equation 1:

$$z_{graph} = Enc_{graph}(\frac{1}{k}\sum_{i=1}^{k} h) \quad (1)$$

The next stage is the Fusion Layer (z), which combines the representation from the text encoder with the graph representation using concatenation to form a single vector. This fusion process can be calculated using Equation 2.

$$z = |h_{text}||z_{graph}| \quad (2)$$

In the final stage, the fused vector is processed through a classification layer to produce the output ($\hat{y}$), which corresponds to one of the three possible labels: True, False, or Not Enough Info. The computation is shown in Equation 3.

$$\hat{y} = Softmax\ (W_z + b) \quad (3)$$

Where W is the weight matrix of the classification layer, and b is the bias vector, which is added after the matrix multiplication W_z to adjust the output. The dimension of b is $b \in \mathbb{R}^3$, corresponding to the number of output classes.

2.2 Dataset

The dataset used in this study is the PolitiFact dataset [24], [25], [26]. The dataset is primarily composed of English-language texts, but some evidence contains multilingual elements through quotations and cross-lingual references. The dataset used in this study consists of 10,000 samples, with an equal distribution of 5,000 labeled as “True” and 5,000 as “False”. The “Not Enough Info” class is not present in the training data, as it is designed to be inferred by the model during prediction. Furthermore, the dataset is divided into 70% for training, 20% for validation, and 10% for testing. Each entry in the dataset consists of three components: Statement (political claim): a single sentence containing the subject’s assertion; Evidence (supporting or refuting information): one or more paragraphs sourced from credible news outlets; Label: the fact verification annotation result (true, false, or not enough info).

An example of the PolitiFact dataset in (JavaScript Object Notation) JSON format can be seen in Figure 2.

{
"Table_ID":"T1"
"Statement”: “John McCain opposed bankruptcy protections for families "who were only in bankruptcy because of medical expenses they couldn't pay.""
"Label”: “true"
"Evidence”: [
0:"Trying to portray his opponent as insensitive to the plight of debt-laden Americans, Sen."
1:"Barack Obama used a June 11, 2008, speech to highlight Sen."
2:"John McCain's support for a 2005 law that made it more difficult for personal bankruptcy filers to escape debts that they could repay."
3:"Specifically, he noted McCain's opposition to an effort to exempt from the law individuals whose medical expenses pushed them into bankruptcy."
4:""John McCain has been part of the problem," Obama said."
]
}

Figure 2. Example of a Claim-Evidence Pair from PolitiFact Dataset

2.3 Preprocessing

In preparing the dataset for model training, several preprocessing steps were applied to maintain consistency and optimize the input quality. All claim and evidence texts were first converted to lowercase, and special characters were normalized using standard

Unicode formatting to avoid inconsistencies. Each sentence was then tokenized using the appropriate tokenizer based on the encoder employed, mBERT, XLM-R, or LaBSE to ensure compatibility with the model's expected input structure.

Redundant evidence linked to the same claim was removed to prevent repetition and reduce potential bias during training. When a combined claim-evidence input exceeded the token limit (typically 512 tokens), the evidence portion was truncated while keeping the claim intact, allowing the model to focus on the main statement. To improve the encoder's understanding of the relationship between claims and evidences, the text was formatted in an instruction-like sequence, similar to a question-answer pair. Unnecessary information such as stopwords, timestamps, and irrelevant news headers was also filtered out. These steps were applied consistently across all data partitions, i.e., training, validation, and testing to ensure uniformity in input processing.

2.4 Training and Evaluation

The model was trained using the AdamW optimizer with a learning rate of 1×10^{-4} [27], and the loss function used was cross-entropy. Regularization was applied using dropout and early stopping based on the F1-score on the validation data. During each training iteration, dropout randomly deactivates a portion of the units (neurons) along with their connections in the network.

Early stopping is an automatic strategy to terminate training if the model's performance no longer improves on the validation set [28], [29]. The training process was conducted over 15 epochs [30] using two NVIDIA T4 GPUs to accelerate parallel computation and multilingual representation processing.

2.5 Evaluation Matrix

The performance evaluation of the DiG-MFV model was conducted using four primary evaluation metrics that are commonly used in multi-class text classification tasks, i.e., accuracy, precision, recall (or sensitivity), and F1-score. These metrics are calculated based on the model's predicted values compared to the actual labels in both the validation and test datasets, by taking into account the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The calculations of these metrics are presented in Equations 4 to 7.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1-Score = 2\,x\,\frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

3. Results and Discussions

Model testing was conducted to compare the performance of DiG-MFV variants that utilize three different multilingual encoder models, i.e., mBERT, XLM-R, and LaBSE. The evaluation was carried out based on accuracy, precision, recall, and F1-score on the validation data.

3.1 Results

The results of this study indicate that all performance indicators for the DiG-MFV model using the LaBSE encoder variant achieved the best performance, with an accuracy of 85.80% and an F1-score of 85.70%. These results surpass both baseline models as well as the other variants within the same architecture. On the other hand, DiG-MFV with mBERT showed moderate performance, while the variant with XLM-R recorded the lowest performance with an F1-score of only 37.54%, despite achieving a relatively high precision score. These findings highlight that the encoder's ability to construct semantic representations of claims and evidences plays a crucial role in the model's effectiveness for political fact verification tasks. The performance results of the DiG-MFV encoder variants and the baseline models are presented in Table 1, indicate that the DiG-MFV LaBSE model has a higher level of accuracy.

Table 1. Performance Comparison of DiG-MFV Variants and Baseline Models

Model	Accuracy	Precision	Recall	F1-Score
DiG-MFV mBERT	65.25%	66.18%	65.25%	64.87%
DiG-MFV XLM-R	50.60%	60.72%	50.60%	36.67%
DiG-MFV LaBSE	85.80%	86.46%	85.80%	85.70%
DGMFP [13]	76.10%	-	-	-
XFEVER mBERT [31]	82.20%	-	-	-
XFEVER XLM-R [31]	85.50%	-	-	-

Another analysis shows that from the early epochs, LaBSE's performance increased steadily until the end of training, reaching 85.80% accuracy, 86.46% precision, 85.80% recall, and an F1-score of 85.70% at epoch 15. This trend reflects LaBSE's ability to form semantic representations that are both convergent and generalizable with respect to claims and evidences. In contrast, mBERT demonstrated moderate performance with gradual improvement across all metrics, although it experienced a slowdown in the final epochs, indicating its limitations in capturing deep semantic relationships between sentences. The analysis at each training epoch can be observed in Figure 3.

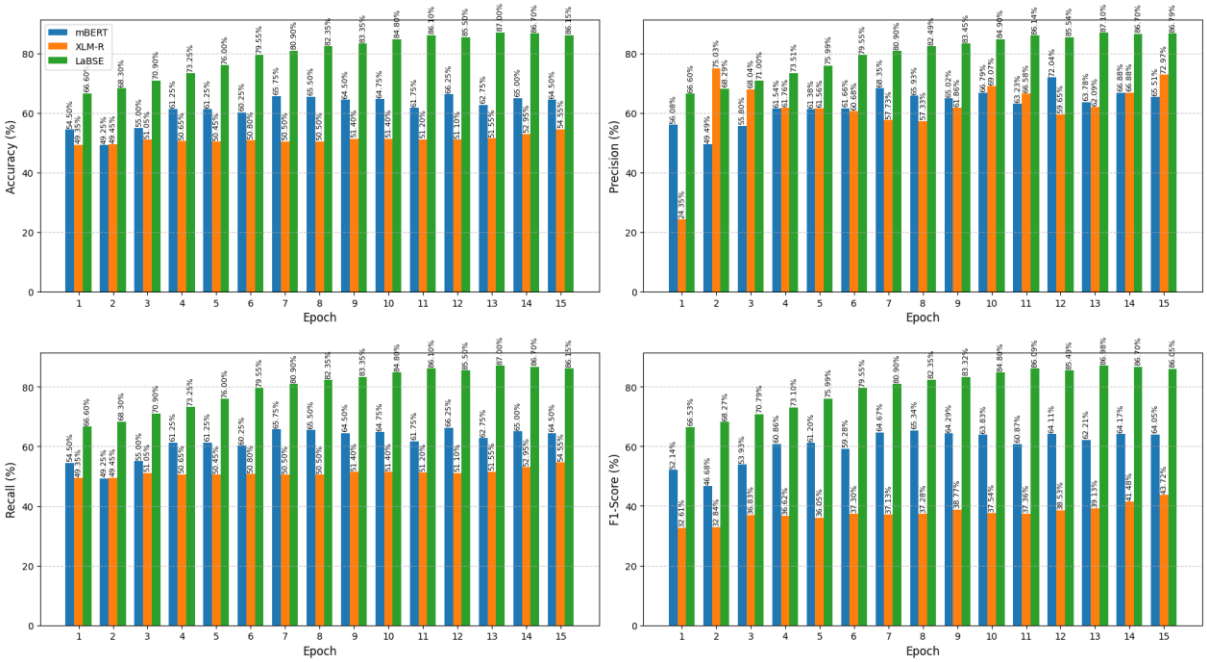

Figure 3. Performance Metrics per Epoch for DiG-MFV with mBERT, XLM-R, and LaBSE Encoders

XLM-R, despite having relatively high precision in the early training stages, exhibited stagnation in recall and F1-score below 60%, suggesting that the model was only able to accurately recognize a small portion of the claims while failing to cover the full distribution of the data.The wide gap between precision and recall in XLM-R also indicates a tendency of the model to overfit to certain subsets of the training data. Furthermore, the analysis of loss per epoch in this study shows that the model using the LaBSE encoder exhibited the most significant and consistent loss reduction throughout all 15 epochs. The visualization of the loss values generated in this study is presented in Figure 4.

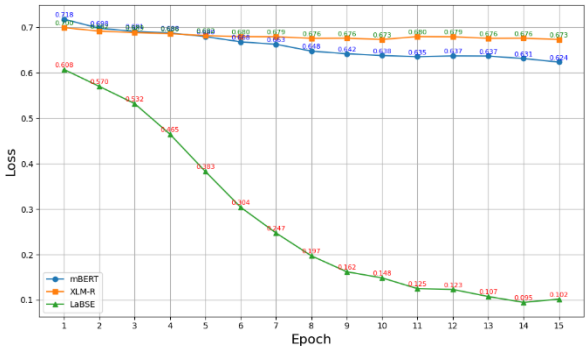

Figure 4. Loss Curves per Epoch for DiG-MFV with mBERT, XLM-R, and LaBSE Encoders

This result indicates that the training process with LaBSE was stable and successfully constructed effective semantic representations of claims and evidences. In contrast, both mBERT and XLM-R showed slower loss reduction and tended to stagnate in the later epochs. The high precision but low recall observed in XLM-R indicates that the model was able to correctly identify only a small portion of the data, but failed to capture the overall pattern comprehensively. Overall, this loss curve reinforces the findings of the study that the choice of encoder has a significant impact on the effectiveness of political fact verification models. LaBSE is proven to be the most optimal encoder for this task, followed by mBERT, while XLM-R requires further improvement or additional strategies to achieve competitive performance.

Finally, compared to its baseline models, the DiG-MFV variant with the LaBSE encoder demonstrates clear advantages over previous approaches, i.e., DGMFP and XFEVER, particularly in terms of architecture for handling multilingual fact verification tasks. DGMFP was specifically designed for table-based facts and heavily relies on symbolic evidence obtained through logical programs, whose construction often depends on semantic parsers and is prone to errors. This approach becomes less relevant when applied to natural language data such as political claims. In contrast, DiG-MFV-LaBSE is designed to work directly with free-form text and integrates two types of graph representations, i.e., the evidence graph and the semantic fusion graph, both constructed from semantic understanding between claims and evidence. Meanwhile, XFEVER is a multilingual benchmark built by translating the FEVER dataset. Although it employs multilingual language models such as mBERT or XLM-R, this approach does not leverage explicit reasoning mechanisms. DiG-MFV-LaBSE combines the strength of multilingual encoders, which are specifically designed for cross-lingual semantic alignment, with a graph-based architecture for cross-evidence reasoning. This combination makes DiG-MFV-LaBSE superior in capturing semantic relationships between sentences

across languages, especially for political claims that often require deep contextual understanding.

3.2 Discussions

The main findings of this study indicate that the choice of encoder significantly impacts the performance of a multi-evidence-based political fact verification system. Among the three encoder variants tested within the DiG-MFV architecture, LaBSE consistently achieved the highest performance across accuracy, precision, recall, and F1-score, outperforming both mBERT and XLM-R. This superiority can be attributed to LaBSE's architecture, which leverages a Siamese Network approach, where two sentences are processed in parallel to generate vector representations that can be directly compared using cosine similarity. LaBSE is specifically optimized for semantic sentence matching tasks through training on a translation ranking objective across more than 100 languages, making it particularly effective for aligning the meaning between claims and evidences in multilingual contexts [22].

In contrast, while mBERT is a widely used multilingual model, it was not explicitly trained to generate sentence-level representations. Instead, it operates as a token-level model trained via masked language modeling (MLM). This limits the stability and consistency of inter-sentence representations in mBERT's embedding space, making it less suitable for classification tasks that rely on semantic relations between sentence pairs. This limitation explains the moderate performance of mBERT in this study [32].

As for XLM-R, despite its strong performance in many multilingual text classification benchmarks, it showed relatively poor performance in political fact verification. One possible reason is that, although XLM-R has a large capacity and is trained on massive multilingual corpora, it is not optimized for cross-sentence alignment, resulting in weaker performance for tasks that require semantic mapping between two distinct texts [33]. Furthermore, the embedding space of XLM-R tends to be scattered, making direct sentence comparison less effective [34].

The loss analysis per epoch supports these findings. The model using LaBSE showed a rapid and consistent decrease in loss, reflecting a stable and effective learning process. Conversely, the loss values for mBERT and XLM-R decreased more slowly and stagnated in the later epochs, indicating difficulty in constructing convergent semantic representations of claim-evidence pairs. Interestingly, although XLM-R achieved relatively high precision, it had low recall and F1-score. This suggests that the model could identify a small subset of patterns with high accuracy but failed to generalize across the broader distribution of claims and evidences, an indication of overfitting on certain segments of the training data.

Although the dual-graph architecture proved effective with LaBSE and provided moderate improvements with mBERT, its integration with XLM-R resulted in low performance, particularly in terms of F1-score. A deeper analysis reveals several possible technical causes. First, XLM-R embeddings tend to be less stable across languages and tasks due to SentencePiece-based tokenization, which leads to inconsistent token segmentation, especially when dealing with domain-specific political vocabulary. This results in misaligned representations between claims and evidence. Second, during training with XLM-R, gradient instability was observed in the early epochs, requiring learning rate adjustments and training restarts, indicating convergence difficulties. Third, the semantic fusion graph is unable to compensate for representational dispersion when the main encoder (XLM-R) fails to generate consistent semantic embeddings. Unlike LaBSE, which is explicitly optimized for sentence-level similarity and produces tightly clustered embeddings, XLM-R generates dispersed and semantically inconsistent representations. As a result, the fusion process between graph and text representations becomes less effective. Moreover, the semantic fusion graph assumes that evidence nodes share latent semantic proximity, an assumption that does not hold when XLM-R embeddings are inconsistent or scattered. These findings indicate that the effectiveness of the graph architecture heavily depends on the quality and consistency of the semantic representations produced by the encoder.

Conceptually, these results reinforce the understanding that political fact verification requires deep and contextual semantic modeling, as political narratives often contain implicit meanings, subtle contradictions, or complex framing. Therefore, an encoder that can precisely capture cross-sentence semantic relationships is essential. LaBSE, with its Siamese architecture and focus on semantic alignment between sentence pairs, proved to be the most effective for this task.

4. Conclusions

The experimental results demonstrate that among the three multilingual encoder variants tested within the DiG-MFV framework, LaBSE consistently delivers the most effective performance. The experimental results demonstrate that the DiG-MFV variant with the LaBSE encoder achieves the best performance, with an accuracy of 85.80% and an F1-score of 85.70%, outperforming other variants and showing competitive results compared to the DGMFP baseline. The superiority of LaBSE lies in its ability to produce stable sentence representations, making it suitable for claim-evidence matching tasks. In contrast, the lower performance of DiG-MFV variants using mBERT and XLM-R indicates that the quality of semantic representations has a significant impact on the accuracy of the reasoning process, even when using the same graph-based architecture. Based on these findings, it can be concluded that combining robust multilingual representations with graph-based reasoning structures

can improve fact verification accuracy. The integration of textual and structural pathways in DiG-MFV offers flexibility in handling various types of political claim data.

Acknowledgements

This research was supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) grant scheme, under Contract No. 126/C3/DT.05.00/PL/2025 (May 28, 2025) and Decree No. 0419/C3/DT.05.00/2025 (May 22, 2025). The project, titled "Pengembangan Seleksi Fitur pada Representasi Graf untuk Penyaringan Bukti dalam Verifikasi Fakta Berita Politik Bahasa Indonesia", was led by Kusrini and implemented through a collaborative agreement between LLDIKTI (Contract No. 0498.21/LL5-INT/AL.04/2025, June 4, 2025) and AMIKOM University (Contract No. 005/KONTRAK-LPPM/AMIKOM/VI/2025, June 5, 2025). We extend our gratitude to all institutions and team members for their invaluable support.

References

[1] A. Rani *et al.*, "FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, May 2023, pp. 10421–10440. doi: 10.18653/v1/2023.acl-long.581.

[2] Z. Chen, F. Zhuang, L. Liao, M. Jia, J. Li, and H. Huang, "Effectively Modeling Sentence Interactions with Factorization Machines for Fact Verification," *IEEE Intell Syst*, vol. 38, no. 5, pp. 18–27, Sep. 2023, doi: 10.1109/MIS.2023.3301170.

[3] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, and A. Bandhakavi, "Logically at Factify 2022: Multimodal Fact Verification," Dec. 2021, doi: https://doi.org/10.48550/arXiv.2112.09253.

[4] N. Agustina, Kusrini, E. Utami, and T. Hidayat, "Systematic Literature Review in the Development of Datasets and Fact Verification Models for Indonesian Language," in *2024 7th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, Sep. 2024, pp. 1–9. doi: 10.1109/IC2IE63342.2024.10748079.

[5] 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)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Sep. 2023, pp. 16190–16206. doi: 10.18653/v1/2023.acl-long.895.

[6] A. Ünver, "Emerging Technologies And Automated Fact-Checking: Tools, Techniques And Algorithms," Aug. 2023. doi: 10.13140/RG.2.2.20514.20165.

[7] A. Athar, S. Ali, M. M. Sheeraz, S. Bhattachariee, and H.-C. Kim, "Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning," no. March, pp. 01–05, 2022, doi: 10.1109/snams53716.2021.9732159.

[8] I. Perikos and S. Souli, "Natural Language Inference with Transformer Ensembles and Explainability Techniques," 2024, doi: 10.3390/electronics.

[9] C. J. Varshney, A. Sharma, and D. P. Yadav, "Sentiment analysis using ensemble classification technique," *2020 IEEE Students' Conference on Engineering and Systems, SCES 2020*, no. July, 2020, doi: 10.1109/SCES50439.2020.9236754.

[10] A. Praseed, J. Rodrigues, and P. S. Thilagam, "Hindi fake news detection using transformer ensembles," *Eng Appl Artif Intell*, vol. 119, Mar. 2023, doi: 10.1016/j.engappai.2022.105731.

[11] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-023-00842-0.

[12] Z. Yang, Y. Xu, J. Hu, and S. Dong, "Generating knowledge aware explanation for natural language inference," *Inf Process Manag*, vol. 60, no. 2, Mar. 2023, doi: 10.1016/j.ipm.2022.103245.

[13] H. Gong, C. Wang, and X. Huang, "Double Graph Attention Network Reasoning Method Based on Filtering and Program-Like Evidence for Table-Based Fact Verification," *IEEE Access*, vol. 11, pp. 86859–86871, 2023, doi: 10.1109/ACCESS.2023.3304915.

[14] C. Liu, Z. Yao, Y. Zhan, X. Ma, S. Pan, and W. Hu, "Gradformer: Graph Transformer with Exponential Decay," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.15729

[15] G. Mitrov, B. Stanoev, S. Gievska, G. Mirceva, and E. Zdravevski, "Combining Semantic Matching, Word Embeddings, Transformers, and LLMs for Enhanced Document Ranking: Application in Systematic Reviews," *Big Data and Cognitive Computing*, vol. 8, no. 9, p. 110, Sep. 2024, doi: 10.3390/bdcc8090110.

[16] L. Wu, D. Yu, P. Liu, C. Gao, and Z. Wang, "Heuristic Heterogeneous Graph Reasoning Networks for Fact Verification," *IEEE Trans Neural Netw Learn Syst*, 2023, doi: 10.1109/TNNLS.2023.3282380.

[17] L. Pan, Y. Zhang, and M.-Y. Kan, "Investigating Zero- and Few-shot Generalization in Fact Verification," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Sep. 2023, pp. 511–524. doi: 10.18653/v1/2023.ijcnlp-main.34.

[18] M. DeHaven and S. Scott, "BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification," in *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Mar. 2023, pp. 58–65. doi: 10.18653/v1/2023.fever-1.6.

[19] M. Naseer, M. Asvial, and R. F. Sari, "An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification," in *3rd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 241–246. doi: 10.1109/ICAIIC51459.2021.9415192.

[20] Z. Liu, C. Xiong, Z. Dai, S. Sun, M. Sun, and Z. Liu, "Adapting Open Domain Fact Extraction and Verification to COVID-FACT through In-Domain Language Modeling," pp. 2395–2400, 2020, doi: 10.18653/v1/2020.findings-emnlp.216.

[21] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[22] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2022, pp. 878–891. doi: 10.18653/v1/2022.acl-long.62.

[23] A. Gaurav, B. B. Gupta, S. Sharma, R. Bansal, and K. T. Chui, "XLM-RoBERTa Based Sentiment Analysis of Tweets on Metaverse and 6G," *Procedia Comput Sci*, vol. 238, pp. 902–907, 2024, doi: 10.1016/j.procs.2024.06.110.

[24] Y. Zhu, J. Si, Y. Zhao, H. Zhu, D. Zhou, and Y. He, "EXPLAIN, EDIT, GENERATE: Rationale-Sensitive Counterfactual Data Augmentation for Multi-hop Fact Verification," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2023, pp. 13377–13392. doi: 10.18653/v1/2023.emnlp-main.826.

[25] V. Bhatnagar, D. Kanojia, and K. Chebrolu, "Harnessing Abstractive Summarization for Fact-Checked Claim Detection," Sep. 2022, Accessed: Jun. 01, 2025. [Online]. Available: https://aclanthology.org/2022.coling-1.259/

[26] Y. Yang, Y. Zhou, Q. Ying, Z. Qian, and X. Zhang, "Search, Examine and Early-Termination: Fake News Detection with Annotation-Free Evidences," Jul. 2024, doi: 10.48550/arXiv.2407.07931.

[27] X. Wang and L. Aitchison, "How to set AdamW's weight decay as you scale model and dataset size," May 2024, doi: https://doi.org/10.48550/arXiv.2405.13698.

[28] F. Ji, X. Zhang, and J. Zhao, "α-EGAN: α-Energy distance GAN with an early stopping rule," *Computer Vision and Image Understanding*, vol. 234, Sep. 2023, doi: 10.1016/j.cviu.2023.103748.

[29] T. Miseta, A. Fodor, and Á. Vathy-Fogarassy, "Surpassing early stopping: A novel correlation-based stopping criterion for neural networks," *Neurocomputing*, vol. 567, Jan. 2024, doi: 10.1016/j.neucom.2023.127028.

[30] B. Goswami, A. B. Somaraj, P. Chakrabarti, R. Gudi, and N. Punjabi, "Classifier Enhanced Deep Learning Model for Erythroblast Differentiation with Limited Data," Nov. 2024, doi: https://doi.org/10.48550/arXiv.2411.15592.

[31] Y.-C. Chang, C. Kruengkrai, and J. Yamagishi, "XFEVER: Exploring Fact Verification across Languages," Oct. 2023, Accessed: Jun. 01, 2025. [Online]. Available: https://aclanthology.org/2023.rocling-1.1/

[32] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2020, pp. 120–130. doi: 10.18653/v1/2020.repl4nlp-1.16.

[33] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[34] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 4512–4525. doi: 10.18653/v1/2020.emnlp-main.365.