ISSN 2723-6471

Comparison of Multilingual Model Sensitivity for Political Fact Verification with Integrated Multi-Evidence

Nova Agustina[1], Kusrini[2,*], Ema Utami[3], Tonny Hidayat[4]

[1,2,3,4]*Department of Informatics Doctorate, Universitas Amikom Yogyakarta, Ring Road Utara Street, Yogyakarta 55281, Indonesia*

(Received: November 1, 2025; Revised: January 1, 2026; Accepted: March 15, 2026; Available online: April 4, 2026)

Abstract

Political news is frequently targeted by fake news on social media, and a key challenge lies in the limited sensitivity of cross-lingual fact verification models to capture semantic relationships between claims and evidence in long-text, multi-evidence settings. This study compares three multilingual Large Language Models, i.e., Multilingual Bidirectional Encoder Representations from Transformers (mBERT), Cross-lingual Language Model-RoBERTa (XLM-R), and Language-Agnostic BERT Sentence Embedding (LaBSE), for political fact verification using an integrated multi-evidence approach. Experiments are conducted on the PolitiFact dataset, with performance evaluated using sensitivity, accuracy, precision, and F1-score metrics. The results indicate that mBERT achieves the highest overall sensitivity at 89.44%, followed by LaBSE at 81.81% and XLM-R at 78.81%. However, mBERT exhibits lower precision, whereas LaBSE provides a better balance between precision (87.02%) and accuracy (86.46%), resulting in an F1-score of 84.33%. XLM-R demonstrates lower sensitivity but maintains competitive precision (85.47%) and accuracy (84.60%), with an F1-score of 82.00%. Sensitivity analysis based on the number of evidence reveals distinct model behaviors, where mBERT achieves its highest observed sensitivity when using six pieces of evidence, XLM-R is more effective under limited evidence conditions, and LaBSE shows a stable and increasing sensitivity trend as the amount of evidence increases. To assess whether these observed performance differences are statistically meaningful, paired statistical significance tests are conducted. The results indicate that the observed performance peak of mBERT at six pieces of evidence does not constitute a statistically dominant global optimum, while XLM-R exhibits a model-specific local optimum under limited evidence conditions and LaBSE demonstrates relatively stable performance across a wide range of evidence sizes. Further statistical analysis shows that XLM-R has the lowest performance variance, while LaBSE exhibits more consistent and statistically robust performance compared to mBERT. Overall, LaBSE is recommended as the most balanced model for multi-evidence-based political fact verification.

Keywords: Political News, Fact Verification, Multi-Evidence, Multilingual Model, Sensitivity Analysis

1. Introduction

Political news often becomes the target of social media, which plays an important role in election campaigns through highly targeted communication strategies [1]. One commonly used approach is fear-based narrative manipulation, such as issues portrayed as "social bombs." In this situation, society is flooded with information, making it difficult to distinguish between valid and fake news. Fake news is designed to spread false, inaccurate, or misleading information, with the deliberate aim of creating public harm or gaining personal benefit. As a result, fake news can trigger public emotions and reinforce existing biases. Its spread in political campaigns can undermine public trust in democratic institutions and the electoral process, posing a significant challenge to the existing information system [2], [3]. The disruption caused by fake news in political contexts is generally expressed through emotional and highly contextual language, which increases cross-lingual linguistic variability and poses significant challenges for computational fact verification systems in capturing semantic relationships between information and supporting evidence.

The issue of fake news can be addressed through various methods, i.e., fact-checking [4], fake news detection [5], [6], and fact verification [7], [8]. Fact verification involves proving claims with supporting or refuting data, making it one of the primary approaches in tackling fake news. However, several studies face challenges in the fact verification modeling process, achieving good and reliable sensitivity results. In this study, sensitivity is defined as the true positive

***Corresponding author: Kusrini (kusrini@amikom.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1198

rate, representing the proportion of supported claims that are correctly identified by the model. Sensitivity is particularly critical in multi-evidence fact verification, as failures to capture relevant claim–evidence relationships may result in missed detections of valid claims. Fact verification analysis related to political news has been previously conducted using the PolitiFact dataset [9]. The limitations of the PolitiFact dataset tend to have low sensitivity to models due to redundant information, posing challenges in determining the relevance and quality of evidence to claims [10]. These limitations are demonstrated by the low average sensitivity of models on the PolitiFact dataset. In a study [11], the average sensitivity was 69.65%, which is 10.15% lower compared to its benchmark dataset, Snopes, which had an average sensitivity of 79.8%. Another study using the PolitiFact dataset also resulted in an average model sensitivity of only 69.03% [12]. Most previous studies have employed sequential model-based approaches to capture interactions between claims and evidence, making the models less capable of capturing semantic relationships between claims and evidence that are separated in long texts. Furthermore, the structure of fact verification datasets, which contain multiple pieces of evidence, has the potential to reduce model sensitivity to the tested dataset. These limitations highlight the need for a more effective approach to filter relevant information and process semantic relationships between claims and multi-evidence.

The currently popular models are Large Language Models (LLMs), which have been used for fact verification in recent years [13], [14]. In text processing cases, LLMs have proven capable of capturing the semantic relationships between claims and evidence. The application of LLMs can be utilized in the Natural Language Processing (NLP) stages by embedding the dataset. The use of LLMs for embedding helps capture linguistic nuances and language context. LLMs are divided into two types: monolingual and multilingual. Multilingual LLMs, which have a broader dimension, have been trained on more than a hundred thousand languages. The advantage of multilingual LLMs lies in their ability to perform transfer learning across languages [15]. Multilingual LLMs that have been applied to text understanding and fact verification tasks include Multilingual Bidirectional Encoder Representations from Transformers, (mBERT) [16], [17], [18], Cross-lingual Language Model-RoBERTa (XLM-R) [19], [20], and Language-Agnostic BERT Sentence Embedding (LaBSE) [21]. However, the application of multilingual LLMs for political news fact verification based on multi-evidence has not yet been explored in depth, presenting a new research opportunity to develop more sensitive and optimal models for understanding the relationship between claims and evidence for the PolitiFact dataset.

Based on the aforementioned description, this study aims to compare the application of multilingual LLMs, i.e., mBERT, XLM-R, and LaBSE, in political news fact verification based on multi-evidence. This research also focuses on optimizing multilingual embeddings to maintain model sensitivity to the relationship between claims and evidence to produce accurate analysis.

2. Literature Review

2.1. Previous Research

Research on automated fact verification in recent years has shown notable progress, particularly through the utilization of deep learning, graph-based reasoning, and multilingual language models. Study [8] marks a shift from evaluating accuracy alone toward analyzing model sensitivity to the number of evidences, demonstrating that approaches based on Graph Attention Networks (GAT) and Kernel Graph Attention Networks (KGAT) achieve optimal performance when the number of evidences is limited (5–6 evidences), but experience sensitivity degradation as the number of evidences increases. These findings indicate that the complexity of relationships among evidences can directly affect model stability in multi-evidence scenarios. A summary comparison of several previous studies relevant to the integration of multilingual models is presented in table 1.

Table 1. Summary Comparison of Related Works

Ref	Background	Method	Result
[8]	This study focuses on evaluating model sensitivity to the number of evidences in fact verification tasks.	GAT and KGAT	The sensitivity results indicate that the model achieves optimal performance when claim–evidence pairs contain 5–6 evidences; however, sensitivity degrades as the number of evidences increases beyond this range.
[22]	This study explores fact verification in a multilingual context with the aim of	BERT	The model is capable of performing cross-lingual fact verification; however, it does not analyze sensitivity to

Ref	Background	Method	Result
	addressing the limitations of monolingual models on cross-lingual datasets.		the number of evidences nor evaluate performance stability in multi-evidence scenarios.
[23]	This study analyzes the extent to which Multilingual BERT (mBERT) truly forms multilingual representations and its capability for zero-shot cross-lingual transfer.	mBERT	The results show that mBERT is able to perform cross-lingual and cross-script transfer, demonstrating its effectiveness in handling multilingual representations, although limitations remain for language pairs that are highly divergent in typological terms.
[24]	This study evaluates the stability of multilingual representations across various NLP tasks.	XLM-R	The model exhibits high cross-lingual stability; however, its sensitivity is lower compared to other multilingual models.
[25]	This study develops cross-lingual semantic embeddings for sentence matching.	LaBSE	The model provides a good balance between precision and recall.
[26]	This study introduces Politi-Fact-Only (PFO), a benchmark dataset derived from PolitiFact by removing post-claim analysis and annotator cues to prevent information leakage in political fact verification evaluation.	RAV (Recon-Answer-Verify)	The evaluation shows an average performance decrease of 11.39% (macro-F1) on PFO compared to the unfiltered version, confirming the presence of leakage bias in the standard PolitiFact dataset. The RAV model outperforms state-of-the-art methods by up to 57.5% macro-F1 on PFO and demonstrates the highest robustness to incomplete evidence.
[27]	This study examines the limitations of one-dimensional truthfulness scales in assessing political misinformation, particularly for claims derived from the PolitiFact dataset, which often exhibit nuances such as bias, incompleteness, or imprecision.	Crowdsourcing-based multidimensional truthfulness assessment with seven dimensions (Correctness, Neutrality, Comprehensibility, Precision, Completeness, Speaker's Trustworthiness, and Informativeness), compared against expert PolitiFact labels.	The results show that multidimensional assessments are reliable and largely independent, exhibit strong correlation with expert PolitiFact labels on the Overall Truthfulness dimension, and provide higher explainability than binary or single-scale classification approaches.

In a cross-lingual context, study [22] extends the scope of fact verification by applying BERT-based models to multilingual datasets, thereby addressing the limitations of monolingual approaches. However, this study does not evaluate sensitivity to the number of evidences nor performance stability in multi-evidence scenarios. Subsequent work [23] strengthens the theoretical foundation for the use of Multilingual BERT (mBERT), demonstrating its ability to produce effective multilingual representations for zero-shot cross-lingual transfer, including across languages with different scripts, although limitations remain for language pairs that are highly divergent in typological terms. Meanwhile, study [24] shows that XLM-R achieves high cross-lingual representation stability, albeit with relatively lower sensitivity compared to other multilingual models. Another multilingual model, LaBSE [25], was developed for cross-lingual semantic matching and has been shown to provide a good balance between precision and recall, making it a strong candidate for claim–evidence matching tasks.

Beyond model-related aspects, dataset quality and truthfulness evaluation have also become key concerns. Study [26] introduces Politi-Fact-Only (PFO) as a benchmark derived from PolitiFact by removing post-claim analysis and annotator cues to prevent information leakage. The evaluation results reveal a significant performance drop compared to the standard PolitiFact dataset, highlighting that many prior models relied on implicit data biases. The Recon–Answer–Verify (RAV) approach proposed in this study is shown to be more robust when dealing with incomplete evidence. Furthermore, study [27] critiques the use of a one-dimensional truthfulness scale in PolitiFact-based datasets and proposes a crowdsourcing-based multidimensional truthfulness assessment. The findings indicate that political claim truthfulness is inherently complex and multidimensional, and therefore cannot be fully captured by binary or single-scale classification.

Overall, the state of the art indicates that although substantial progress has been made in graph-based multi-evidence reasoning, multilingual models, and PolitiFact-based benchmarks, a research gap remains, particularly in the analysis of sensitivity and stability of multilingual models with respect to variations in the number of evidences in the political domain. This gap constitutes the primary motivation for the present study.

3. Methodology

This study employs an experimental design to evaluate the performance of multilingual models in integrated multi-evidence political news fact verification. The study aims to optimize the fact verification process by applying multilingual NLP-based models, namely mBERT, XLM-R, and LaBSE, to capture semantic relationships between

claims and supporting evidence across languages. The evaluation process assesses model performance on multi-evidence inputs by measuring sensitivity and other standard classification metrics. This experiment is designed to compare the effectiveness of multilingual models in handling cross-lingual and multi-evidence fact verification tasks. In this study, sensitivity is defined as the true positive rate, representing the proportion of factually supported claims that are correctly identified by the model. Sensitivity serves as a key evaluation metric for assessing a model's effectiveness in capturing relevant semantic relationships between claims and supporting evidence in cross-lingual and multi-evidence settings. An illustration of the research stages used in this study can be seen in figure 1.

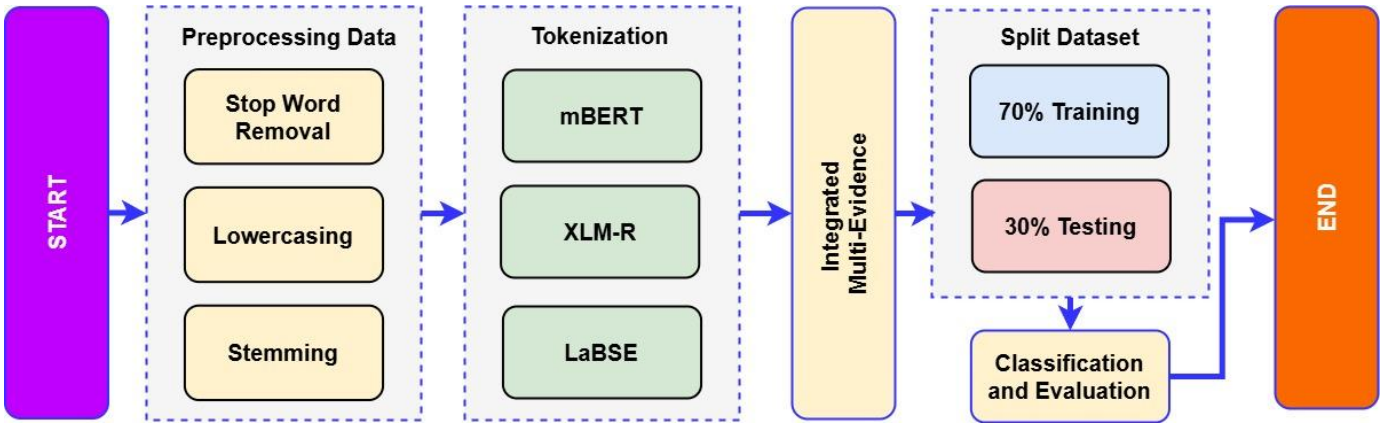


Figure 1. Research Workflow for Political News Fact Verification Using Multilingual Models (mBERT, XLM-R, and LaBSE)

3.1. Preprocessing Data

This study leverages the PolitiFact dataset, sourced from PolitiFact on Papers with Code as the primary resource for testing and evaluating multi-evidence-based fact verification models. The dataset consists of 10,000 data, containing political claims that have been verified by fact-checking teams, accompanied by truth labels and relevant supporting evidence. This dataset selected due to its data structure, which supports multi-evidence analysis and has been widely used in similar research to evaluate the effectiveness of fact verification models. The structure of the PolitiFact dataset used in this study is summarized in table 2.

Table 2. Structure of PolitiFact Dataset

Element	Description
Table_Id	Unique identifier for the fact verification instance.
Statement	The political claim that needs verification.
Label	The factual evaluation of the claim, categorizing its truthfulness.
Evidence	A list of supporting textual evidence used to verify the claim.

Furthermore, before being used in the experiment, the dataset undergoes a pre-processing stage to ensure data consistency and quality. The applied pre-processing stages include:

3.1.1. Stop Word Removal

In this study, the stop word removal process conducted using the NLTK (Natural Language Toolkit) library. Stop words are common words that frequently appear in text but do not contribute significantly to contextual understanding, such as "for", "who", "only", "in", and "because". Removing stop words makes the text more concise and focuses on the words that are crucial for analysis. For example, in the dataset, evidence sentences such as "John McCain opposed bankruptcy protections for families, who were only in bankruptcy because of medical expenses they couldn't pay" After applying stop word removal, the sentence becomes "John McCain opposed bankruptcy protections families, bankruptcy medical expenses couldn't pay"

3.1.2. Lowercasing

Lowercasing is the process of converting all letters in the text to lowercase to maintain data consistency. This process is crucial in Natural Language Processing (NLP) as models are often case-sensitive, meaning uppercase and lowercase letters can be treated differently. For example, the sentence obtained after stop word removal, "John McCain opposed bankruptcy protections families, bankruptcy medical expenses couldn't pay," becomes "john mccain opposed bankruptcy protections families, bankruptcy medical expenses couldn't pay" after applying lowercasing.

3.1.3. Stemming

Stemming is the process of converting words to their root forms by trimming prefixes and suffixes. This method aims to simplify the text so that different word forms can be treated as a single entity. The stemming technique used in this study is from NLTK, such as PorterStemmer or SnowballStemmer. For example, the sentence obtained after lowercasing, "john mccain opposed bankruptcy protections families, bankruptcy medical expenses couldn't pay," becomes "john mccain oppos bankrupt protect famili, bankrupt medic expens could not pay" after applying stemming. The stemming process is applied at the initial normalization stage to reduce surface-level word variation before the tokenization process is performed.

3.2. Tokenization

Tokenization in this study utilizes the built-in tokenizers of the selected multilingual Large Language Models (LLMs), i.e., Multilingual BERT (mBERT), XLM-RoBERTa (XLM-R), and Language-Agnostic BERT Sentence Embedding (LaBSE). In general, these three methods share the same tokenization technique, which involves breaking down text into smaller units, such as words or subwords (subword tokenization), aimed at handling language variations and enriching the semantic representation of text. This technique allows the model to handle rare words by splitting complex words into more common subwords, thereby reducing the Out-Of-Vocabulary (OOV) issue and enhancing the understanding of different language contexts. All three methods employ subword-based tokenization approaches:

3.2.1. Multilingual BERT (mBERT)

The mBERT model is trained using 104 languages with dataset sources from Wikipedia. The tokenization method used in mBERT is WordPiece, which segments words based on the most common subword frequencies. The tokenization formula for the mBERT model can be seen in Equation 1.

$$T(S) = \arg \max_{\{t_1,t_2,\ldots,t_k\}} \prod_{i=1}^{k} P(t_i|t_1,t_2,\ldots,t_{i-1}) \quad (1)$$

Where Tokenization (T) is performed by assigning tokens (t) to the sentence (S) and finding the combination of tokens t_i with the highest probability based on the word distribution in the vocabulary (k), where k denotes the number of subword tokens generated during the WordPiece tokenization process, and each token t_i is selected to maximize the likelihood given the preceding subword sequence. The most frequently occurring subword (arg max) is prioritized in the segmentation process.

3.2.2. XLM-RoBERTa (XLM-R)

The XLM-R model is trained using 100 languages with dataset sources from CommonCrawl. The tokenization method used in XLM-R is SentencePiece. The tokenization formula for the XLM-R model can be seen in Equation 2.

$$T(S) = \arg \max_{\{s_1,s_2,\ldots,s_m\}} P(s_i|s_1,s_2,\ldots,s_m|S) \quad (2)$$

Where Tokenization (T) is performed by assigning subwords generated based on the maximum probability from the unigram model (s) to the sentence (S), where s_i represents subword units generated by the SentencePiece unigram language model, which assigns probabilities to candidate segmentations of the input sentence.

3.2.3. Language-Agnostic BERT Sentence Embedding (LaBSE)

The LaBSE model is trained using 109 languages with dataset sources from the Multilingual Dataset. The tokenization method used in LaBSE is the same as mBERT, which utilizes WordPiece but with specific optimizations for cross-lingual semantic matching. The tokenization formula for the LaBSE model can be seen in Equation 3.

 ISSN 2723-6471

$$T(S) = \{t_1, t_2, \dots, t_n\} \text{ with } t_n \in \text{V} \quad (3)$$

Where Tokenization (T) is performed by assigning tokens (t) to the sentence (S) and finding the combination of tokens t_i, where t represents the vocabulary generated from WordPiece (V).

3.3. Integrated Multi-Evidence (Reconstruction Tokenization)

The Integrated multi-evidence approach in this study is used in the fact verification process by combining multiple pieces of evidence into a single structured representation in the PolitiFact dataset. After the tokenization process, a reconstruction step is performed to ensure that the claims and evidence can be accurately mapped back to the original text. The integrated multi-evidence procedure for fact verification based on claims and multiple pieces of evidence. The process begins by receiving several key inputs, i.e., the set of claims (C), the set of evidence documents (E), the pre-trained tokenizer (T), the pre-trained multilingual model (M), and the maximum token length (Lmax), which is set to 512, as it represents the maximum limit that handled by pre-trained language models such as mBERT, XLM-R, and LaBSE. This process involves the procedure summarized in algorithm 1.

Algorithm 1. Pseudocode of Integrated Multi-Evidence

Input
C = set of claims
E = set of evidence documents ($E_1, E_2, E_3, \dots E_n$)
T = pre-trained tokenizer
M = pre-trained multilingual model
L_{max} = maximum token length

Output
Integrated representation of claims with multi-evidence

Algorithm
for each claim C_i in C do
    $C_{processed}$ ← preprocess (C_i)
    for each evidence E_j in E do
        $E_{processed}$ ← preprocess (E_j)
    end

    $Combined_{text}$ ← concatenate ($C_{processed}$, $E_{processed}$, "[SEP]")
    If (length ($Combined_{text}$) > L_{max} then
        $Combined_{text}$ ← truncate ($Combined_{text}$, L_{max})
    end

    $Tokenized_{inputs}$ ← T ($Combined_{text}$)
    $Embedded_{representation}$ ← M ($Tokenized_{inputs}$)
    Store $Embedded_{representation}$ in dataset
end

In Algorithm 1, The algorithm iterates through each claim in the claim set (C). Each claim undergoes a preprocessing stage first, which includes steps such as stopword removal, lowercasing, and stemming on both C and the evidence set (E). After the claims and evidence are processed, the next step is to concatenate the processed claim with the processed evidence using a special separator token "[SEP]". The "[SEP]" token functions to explicitly mark the boundary between the claim and each piece of evidence, enabling the model to distinguish and model their semantic relationships. In the process of integrating claims and multiple pieces of evidence using the special "[SEP]" token, this study employs attention masks that are automatically generated by the native tokenizer of each multilingual model to distinguish valid tokens from padding. Meanwhile, positional encoding follows the default scheme of the underlying model architecture. Accordingly, the "[SEP]" token functions solely as a segment boundary marker and does not require any manual adjustment to the attention mask or positional encoding mechanisms.

Furthermore, if the combined text length exceeds the maximum limit set in Lmax, truncation is performed to ensure that the data fed into the model remains within the model's processing capabilities. Once the concatenation and truncation processes are complete, the resulting text undergoes tokenization using the pre-trained tokenizer (T). The

 ISSN 2723-6471

tokenized output is then passed to the pre-trained multilingual model (M) to generate an Embeddedrepresentation of the processed claims and evidence. The final stage of this algorithm is storing the generated representations into the dataset for further training or evaluation purposes in the fact verification task. This process aims to enable the model to understand the relationship between claims and evidence. Although no additional or customized attention mechanisms were introduced, the models inherently employ self-attention through the transformer architecture. In this study, the attention_mask generated during tokenization utilized to control token-level attention by masking padded positions, ensuring that only valid claim and evidence tokens contribute to the self-attention computation. No explicit attention masking or positional encoding adaptations were applied beyond the default configuration of the pretrained models. An illustration of the multi-evidence integration in this study can be seen in figure 2.

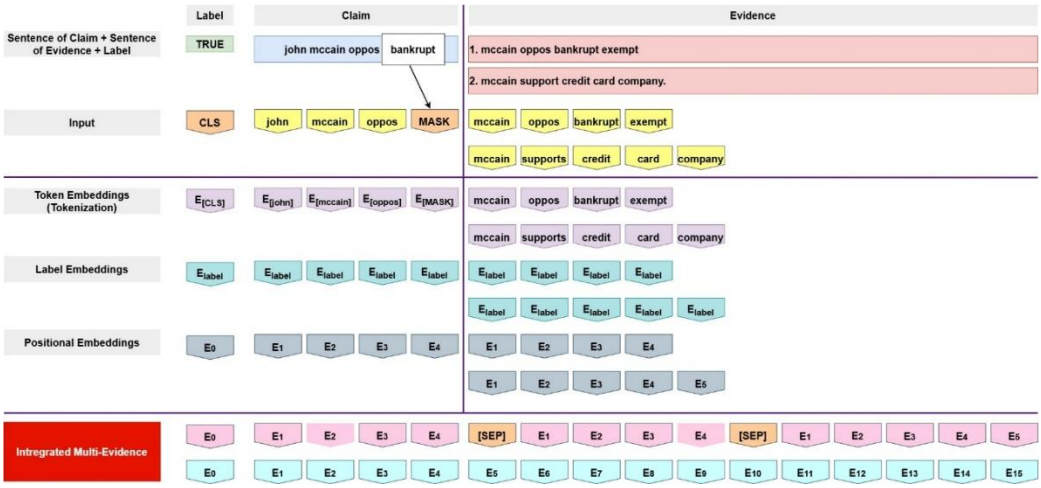


Figure 2. Illustration of Integrated Multi-Evidence

In figure 2, the claim and supporting evidence are combined into a single format using a special [SEP] token as a separator between the claim and each available piece of evidence. Technically, the formula used to construct the Integrated multi-evidence can be seen in Equation 4.

$$IME(C,\{E_i\}_{i=1}^{n}) = T\left([CLS] \oplus C \oplus \bigoplus_{i=1}^{n} ([SEP] \oplus E_i)\right) \quad (4)$$

Where IME represents Integrated multi-evidence, C refers to the claim text, $\{E_i\}_{i=1}^{n}$ represents a set of evidence sentences, CLS and SEP is the separator token used to distinguish between the claim and evidence, $\oplus$ denotes sequence concatenation, and T(.) denotes the tokenizer defined in Equations (1)–(3).

3.4. Split Dataset

In this study, the dataset is divided into two parts with a 70-30 ratio to ensure an effective balance between training and testing data [28]. Specifically, 70% of the dataset is allocated for training purposes, allowing the model to learn patterns and relationships between claims and evidence, while the remaining 30% is used for testing to evaluate the model's performance and generalization ability. The dataset splitting process is performed using random sampling with a fixed random seed to ensure reproducibility of the experimental results. This division ensures that sufficient data are available for both model training and performance evaluation on unseen political news claims.

3.5. Classification and Evaluation

In this study, the classification process is carried out by training and testing the mBERT, XLM-R, and LaBSE models to verify political fact claims using the PolitiFact dataset. These models classify each claim into predefined categories based on its truthfulness level, utilizing the Integrated Multi-Evidence approach to analyze the relationship between claims and supporting evidence. The models generate probability scores for each class, and the final prediction is based on the highest probability. The success of the classification is measured by evaluating the extent to which the models can distinguish between true and false claims. Sensitivity is utilized to evaluate the model's performance. The sensitivity measurements for the mBERT, XLM-R, and LaBSE models are derived from the evaluation of algorithm performance based on the indicators True Positive (TP), True Negative (TN), False Positive (FP), and False Negative

(FN). Sensitivity serves as an indicator of the model's ability to correctly identify positive instances [29]. Sensitivity is calculated using Equation 5.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

Furthermore, as an evaluation consideration, this study employs additional indicators, i.e., accuracy, precision, and F1-score [30]. Accuracy is used to measure the overall proportion of correct predictions compared to the total observations. Precision is used to evaluate the extent to which the model can correctly identify true claims from all claims predicted as true. F1-score is used to provide a balance between precision and sensitivity in assessing the overall performance of the model. The calculations for accuracy, precision, and F1-score can be determined using Equations 6 to 8.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - Score = \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{8}$$

4. Results and Discussion

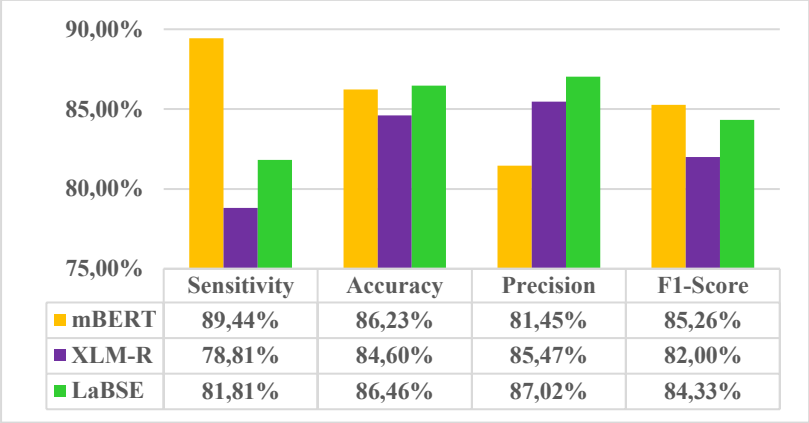
This section presents the evaluation results of multilingual models used in political news fact verification, focusing on the effectiveness of the integrated multi-evidence approach. The discussion includes performance comparisons among the models as well as an analysis of their strengths and limitations in handling multi-evidence claims within the PolitiFact dataset. This analysis aims to highlight the impact of different tokenization techniques and the overall performance of each model based on sensitivity, accuracy, precision, and F1-score metrics. In this study, we used 10 epochs, a learning rate of 0.01, and a batch size of 8 to train multilingual models for political news fact verification. Additionally, the evaluation is conducted under two different scenarios: testing with all evidence and testing with a limited number of evidence. In the full evidence testing scenario, all available evidence for each claim is used in the verification process. Furthermore, under the limited evidence scenario, this study tested the model's sensitivity by restricting the number of evidence provided from 1 to 7, as done in the study [8]. This approach aims to analyze the model's performance when exposed to varying amounts of evidence and to determine its adaptability and robustness in different verification scenarios.

Moreover, statistical analysis is conducted to evaluate the model's consistency and reliability across these conditions. The methods used included calculating standard deviation, t-statistics, and p-values to determine the significance of performance differences among the models. This analysis aims to ensure that the results obtained are statistically valid and provide deeper insights into the strengths and weaknesses of each model in handling political fact verification tasks based on multi-evidence.

4.1. Results

The performance evaluation of the mBERT, XLM-R, and LaBSE models in the task of political news fact verification using the PolitiFact dataset shows variations in sensitivity, accuracy, precision, and F1-score metrics. The evaluation results indicate that the mBERT model achieved the highest sensitivity at 89.44%, followed by LaBSE with 81.81%, and XLM-R with the lowest score of 78.81%. The high sensitivity of mBERT demonstrates that this model can effectively identify true claims and capture the relationship between claims and evidence. On the other hand, XLM-R exhibits the lowest sensitivity, indicating difficulties in capturing the relationship between claims and evidence. The evaluation results in figure 3 show that each model has strengths and limitations in handling claims and evidence using the Integrated multi-evidence approach, which combines multiple pieces of evidence into a single structured representation. The comparative performance evaluation results of the mBERT, XLM-R, and LaBSE models are presented in figure 3.

ISSN 2723-6471

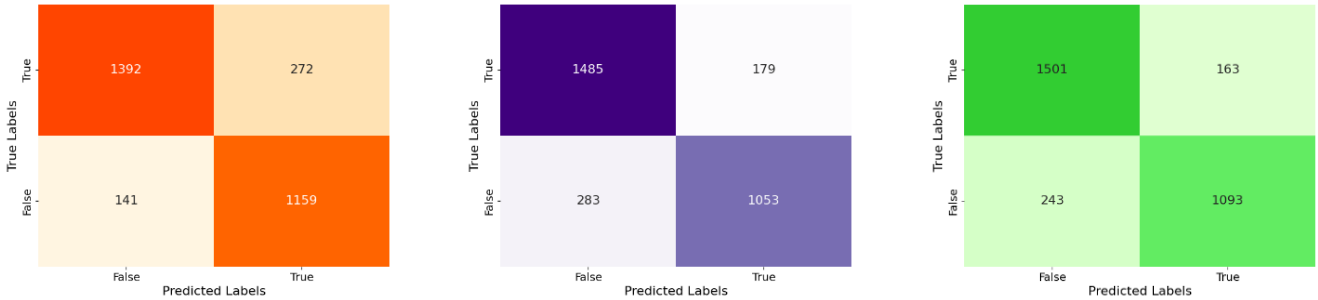


	Sensitivity	Accuracy	Precision	F1-Score
mBERT	89,44%	86,23%	81,45%	85,26%
XLM-R	78,81%	84,60%	85,47%	82,00%
LaBSE	81,81%	86,46%	87,02%	84,33%

Figure 3. Comparison of Sensitivity, Accuracy, Precision, and F1-Score of mBERT, XLM-R, and LaBSE on the PolitiFact Dataset

Furthermore, LaBSE achieved the highest accuracy with a score of 86.46%, surpassing the other models, followed by mBERT with 86.23%, and XLM-R with 84.60%. The slightly higher accuracy of LaBSE demonstrates its superior ability to handle semantic similarity. LaBSE is specifically designed to support semantic matching tasks, making it advantageous in dealing with variations in word and phrase usage within political claims. XLM-R exhibited lower accuracy due to challenges in handling noise within the PolitiFact dataset, such as claims with ambiguous language formulations and evidence that may not explicitly support the claims. This lower accuracy indicates that XLM-R tends to produce more errors when classifying complex claims. In the evaluation of precision measurement, LaBSE achieved the highest precision at 87.02%, followed by XLM-R with 85.47%, and mBERT with the lowest score of 81.45%. LaBSE's superiority in precision indicates that it performs better in making positive decisions by minimizing errors in claims classified as true.

The lower precision performance of mBERT suggests that this model tends to be more permissive in classifying claims as true, even in cases where the available evidence may not be sufficiently strong. This occurs because mBERT focuses more on sensitivity, aiming to capture as many relevant claims as possible. The next performance metric is the F1-score, which shows that LaBSE achieved an F1-score of 84.33%, slightly higher than mBERT at 85.26%, while XLM-R scored 82.00%. Although mBERT has higher sensitivity, its lower F1-score compared to LaBSE indicates a trade-off between sensitivity and precision. Furthermore, the results of the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts for each model can be seen in figure 4.



(a) mBERT (b) XLM-R (c) LaBSE

Figure 4. Confusion Matrices of mBERT, XLM-R, and LaBSE for Political Fact Verification.

Based on the obtained confusion matrices, each model exhibits distinct classification error characteristics. The mBERT model records a high number of True Positives (TP = 1392), indicating a strong capability to broadly identify supported claims. However, mBERT also shows a relatively high number of False Positives (FP = 272), suggesting a tendency to classify claims as supported even when the available evidence does not fully justify such predictions. This pattern is consistent with mBERT's high sensitivity, while also explaining the reduction in precision due to an increased number of positive misclassifications.The XLM-R model produces fewer false positives (FP = 179) compared to mBERT but exhibits a higher number of False Negatives (FN = 283).

 ISSN 2723-6471

This result indicates that XLM-R adopts a more conservative prediction strategy, reducing positive misclassifications at the cost of an increased risk of failing to detect supported claims. Such behavior reflects the trade-off between sensitivity and performance stability.In contrast, the LaBSE model achieves the highest number of True Positives (TP = 1501) while maintaining the lowest number of False Positives (FP = 163) among the evaluated models. This combination demonstrates LaBSE's ability to sustain a high detection rate of supported claims while effectively limiting positive classification errors. Furthermore, LaBSE records a high number of True Negatives (TN = 1093), indicating a stronger capability to consistently filter unsupported claims. Although False Negatives (FN = 243) are still present, their number is more balanced compared to XLM-R, resulting in a more stable trade-off between sensitivity and precision. Furthermore, the model sensitivity testing against the number of evidence is summarized in figure 5.

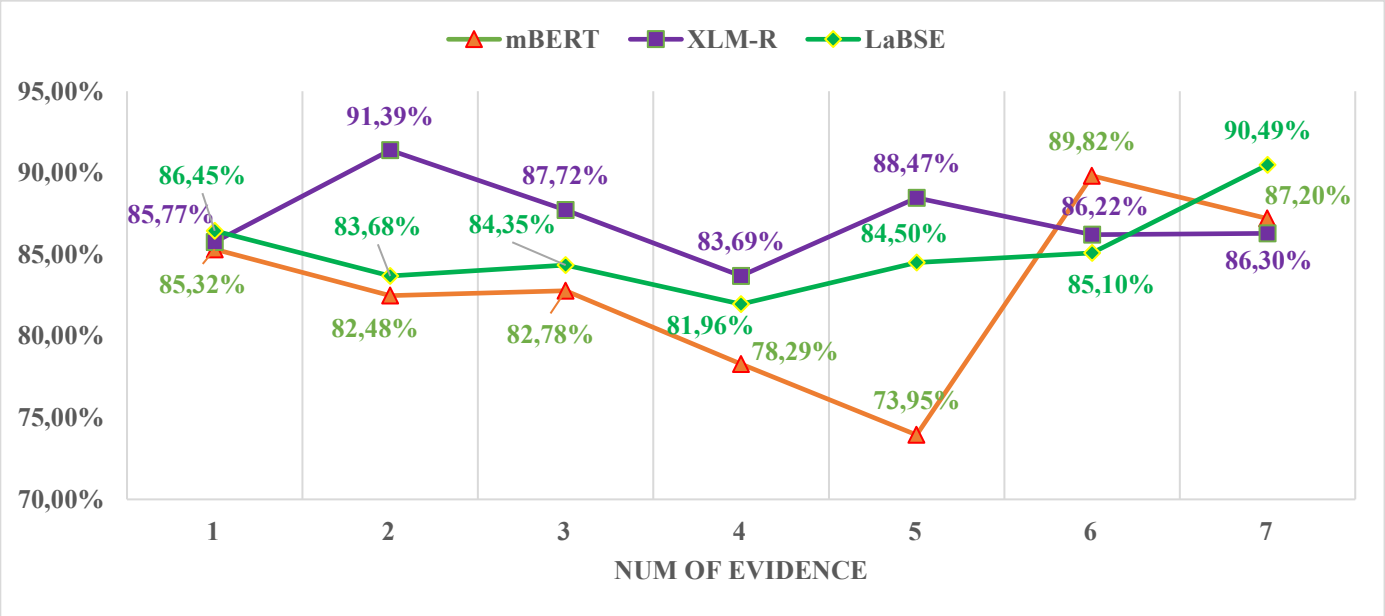


Figure 5. Sensitivity Analysis of mBERT, XLM-R, and LaBSE Based on the Number of Evidence

In figure 5, the models exhibit distinct performance characteristics in handling claims with varying numbers of evidence. The mBERT model achieves its highest observed sensitivity at the sixth piece of evidence, reaching 89.82%, while lower sensitivity values are observed when fewer evidence items are used (K = 1–5). This indicates that an increase in the number of evidence items tends to correlate with improved sensitivity for mBERT, although performance variability remains across different values of K.The XLM-R model demonstrates a different pattern, with its highest observed sensitivity occurring at the second piece of evidence (91.39%). Beyond this point, sensitivity tends to decline and reaches its lowest value of 78.29% at higher evidence counts. Although XLM-R maintains relatively stable performance at subsequent K values, it does not return to the peak sensitivity level observed at K = 2. In contrast, the LaBSE model shows a more stable and gradual increase in sensitivity as the number of evidence items increases, achieving its highest observed sensitivity at the seventh piece of evidence with a value of 90.49%. These findings indicate that LaBSE is able to maintain relatively consistent performance across different evidence configurations, particularly in scenarios involving a larger number of evidence sources.

The results of each model's performance indicators demonstrate their respective strengths and limitations. Therefore, statistical testing is conducted to determine the significance of performance differences among the three models. The statistical analysis is performed to evaluate whether there are significant differences in performance evaluation results, which include calculating the standard deviation to measure data variability, the t-statistic to test the difference in average performance across models, and the p-value to determine the statistical significance of these differences. In this study, statistical hypothesis testing is employed to evaluate whether the observed performance differences among models are statistically significant. The null hypothesis (H_0) assumes that there is no significant performance difference between the compared models, while the alternative hypothesis (H_1) assumes that a performance difference exists at a significance level of $\alpha = 0.05$. The statistical test results are presented in table 3, which provides information on the

ISSN 2723-6471

consistency and reliability of each model's performance in the multi-evidence-based political news fact verification task.

Table 3. Statistical Analysis of mBERT, XLM-R, and LaBSE Based on Standard Deviation, t-Statistic, and p-Value.

Model	Standard Deviation	t-statistic	p-value
mBERT	0.4998	-3.392	0.0006
XLM-R	0.4919	2.714	0.0006
LaBSE	0.4933	2.085	0.0370

Based on the performance evaluation results of the mBERT, XLM-R, and LaBSE models in the political fact verification task using the PolitiFact dataset, significant performance variations were observed among the three models. In terms of sensitivity, mBERT achieved the highest performance with a score of 89.44%, followed by LaBSE at 81.81%, and XLM-R with the lowest score of 78.81%. These results indicate that mBERT achieves higher sensitivity in detecting true claims compared to the other models. However, the statistical analysis in table 3 shows that mBERT has a higher standard deviation (0.4998), indicating greater variability in performance outcomes compared to XLM-R and LaBSE. On the other hand, XLM-R has the lowest standard deviation of 0.4919, demonstrating better performance consistency compared to the other models, despite its lower sensitivity. Additionally, the p-values obtained for mBERT and XLM-R (both at 0.0006) indicate that the observed performance differences among the models are statistically significant at the 0.05 significance level. Meanwhile, LaBSE has a p-value of 0.037, which also indicates statistically significant differences, albeit with a weaker level of significance compared to mBERT and XLM-R.

4.2. Statistical Significance Analysis of Evidence Size (K)

In this study, the evaluation aims to determine whether prior claims stating that mBERT and LaBSE achieve optimal performance with six pieces of evidence (K = 6), and XLM-R with two pieces of evidence (K = 2), are statistically supported. To this end, paired significance analysis is conducted using the McNemar test on the same test dataset. This test is employed to assess whether the performance differences between two evidence-size configurations reflect statistically significant differences or are merely attributable to random variation. In this testing framework, the null hypothesis (H_0) states that there is no significant difference in classification performance between the compared evidence-size configurations, while the alternative hypothesis (H_1) states that a significant performance difference exists. The significance level is set at $\alpha = 0.05$, and all p-values are corrected using the Holm method to control for errors arising from multiple comparisons. The results of the significance tests are reported in table 4.

Table 4. Paired McNemar Test Results for Evidence Size Selection Across Multilingual Models

Model	Comparison	p (McNemar)	p (Holm)	Significant
mBERT	6 vs 1	0.963	1.000	No
	6 vs 2	0.640	1.000	No
	6 vs 3	0	0	True
	6 vs 4	0.299	1.000	No
	6 vs 5	0	0	True
	6 vs 7	0.658	1.000	No
XLM-R	2 vs 1	1.000	1.000	No
	2 vs 3	0	0	Yes
	2 vs 4	0	0	Yes
	2 vs 5	0	0	Yes
	2 vs 6	1.000	1.000	No
	2 vs 7	0	0	Yes
LaBSE	6 vs 1	0.153	0.921	No
	6 vs 2	0.00725	0.058	No
	6 vs 3	0.709	1.000	No

Model	Comparison	p (McNemar)	p (Holm)	Significant
	6 vs 4	0.692	1.000	No
	6 vs 5	1	1	No
	6 vs 7	0	0	Yes

The statistical analysis results indicated that the mBERT model with configuration K = 6 exhibits statistically significant differences when compared with K = 3 and K = 5. However, no significant differences are observed when K = 6 is compared with K = 1, K = 2, K = 4, and K = 7. These findings suggest that although K = 6 yields better performance than certain configurations, it cannot be considered a statistically dominant global optimum. For the XLM-R model, a different pattern emerges. The McNemar test results show that K = 2 acts as a local reference point, exhibiting statistically significant differences relative to K = 3, K = 4, K = 5, and K = 7, but no significant differences when compared with K = 1 and K = 6. This indicates that the optimal evidence size is model-dependent rather than universal across multilingual models. In contrast, for the LaBSE model, the statistical significance analysis reveals that configuration K = 6 differs significantly only when compared with K = 7. No significant differences are observed between K = 6 and the other K configurations. These findings indicate that LaBSE's performance is relatively stable with respect to variations in the number of evidence items within the range K = 1 to K = 6, and does not exhibit a strong, statistically significant performance peak at K = 6.

Overall, the statistical significance analysis indicates that there is no single evidence size (K) that can be considered universally optimal for all evaluated models. For mBERT, the K = 6 configuration shows significant advantages only over certain configurations and does not form a global optimum. For XLM-R, the most statistically significant K value occurs at K = 2, whereas for LaBSE, performance remains relatively stable across the range K = 1 to K = 6 without exhibiting a statistically significant performance peak. These findings confirm that the choice of evidence size is model-dependent, and assigning a single K value as optimal for all multilingual models is not supported by consistent statistical evidence.

4.3. Discussion

The results of this study highlight the various strengths and limitations of multilingual models, i.e., mBERT, XLM-R, and LaBSE in verifying political news facts using the PolitiFact dataset. A comparison of these models has been conducted in previous studies; however, it is used to calculate text correlation coefficients [31]. In this study, these models successfully demonstrated high sensitivity for political news fact verification. Each model demonstrates unique performance characteristics across different evaluation metrics, providing valuable insights into their application in multi-evidence-based fact verification tasks. The integrated multi-evidence approach implemented in this study aims to enhance the models' ability to understand the relationships between claims and multiple pieces of evidence. One of the findings of this study is that the integrated multi-evidence approaches successfully improved the F1-Score, with each model achieving a score above 80%. This result is higher compared to other studies, which generally obtained an F1-Score of around 70% or even lower [10], [11], [12]. This study consolidating multiple pieces of evidence into a single structured representation using the "[SEP]" separator token, this approach allows the models to analyze the relationships between claims and various sources of evidence more effectively. Based on the evaluation results, the integrated multi-evidence approach offers significant benefits in improving model sensitivity, particularly for mBERT, which achieved the highest performance in detecting true claims.

However, challenges arise in terms of precision, as the model tends to be more permissive in accepting claims that are not sufficiently supported by relevant evidence. LaBSE demonstrates a better balance in leveraging information from various pieces of evidence with higher precision, indicating its ability to filter out irrelevant information effectively. On the other hand, the integrated multi-evidence approach faces challenges in XLM-R when handling claims with multiple pieces of evidence that have varying linguistic structures. XLM-R, with its SentencePiece tokenization, struggles to capture the semantic relationships between different pieces of evidence, resulting in lower sensitivity compared to mBERT and LaBSE. This finding suggests that the integrated multi-evidence approach requires further optimization in processing more complex evidence contexts. The implementation of the integrated multi-evidence approach for the mBERT model can be accessed at the following https://www.kaggle.com/code/novaagustina/mbert-politifact. For example, the results of political news fact verification can be seen in figure 6.

 ISSN 2723-6471

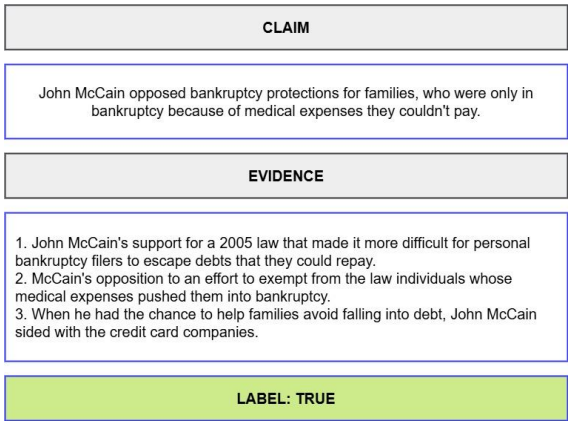


Figure 6. Example of political news fact verification showing a claim and supporting evidence, with the final classification

One of the other findings of this study is the trade-off between sensitivity and precision in the evaluated models. The mBERT model demonstrated the highest sensitivity, indicating its strong capability in detecting true claims and effectively capturing relationships between claims and evidence. However, this high sensitivity comes at the expense of precision, as mBERT tends to classify more claims as true, potentially leading to an increase in false positives. This characteristic makes mBERT more suitable for applications that prioritize sensitivity, such as early misinformation detection, but less ideal for decision-making processes that require high precision. However, the LaBSE model exhibits a balance between sensitivity and precision. Its relatively high precision indicates that the model can better minimize false positives, making it a more reliable choice for applications that require accurate fact verification. LaBSE's ability to maintain this balance can be attributed to its superior capability in matching claims with evidence compared to the other models. The implementation of the integrated multi-evidence approach for the XLM-R model can be accessed at the following link: https://www.kaggle.com/code/novaagustina/xlm-r-politifact.

In addition to accuracy and precision, the consistency and reliability of model performance were evaluated through statistical testing. The results indicate that the XLM-R model has the lowest performance variance, demonstrating that it is more stable and consistent in classifying claims and evidence. This stability is an essential factor for systems that require uniform performance across various scenarios. However, despite its consistency, XLM-R has lower sensitivity, which may hinder its ability to handle complex claims effectively. Furthermore, mBERT exhibits higher performance variability, meaning that its generalization capability is inconsistent across different claim scenarios. This variability can be attributed to the model's reliance on specific linguistic structures present in its training data, making it challenging to handle more ambiguous claims effectively. The implementation of the integrated multi-evidence approach for the LaBSE model can be accessed at the following link: https://www.kaggle.com/code/novaagustina/labse-politifact.

To address the observed trade-off between sensitivity and precision, several technical strategies can be considered as directions for future research. One approach involves implementing evidence selection or weighting mechanisms to reduce the influence of redundant or less relevant information, which may otherwise increase the number of false positives. Additionally, the use of adaptive decision thresholds at the classification stage can help balance sensitivity and precision according to specific application contexts. Another promising direction is the integration of attention mechanisms or confidence-based scoring schemes to emphasize the most informative evidence. These strategies have the potential to mitigate the sensitivity–precision trade-off without sacrificing the benefits of an integrated multi-evidence approach.

5. Conclusion

This study evaluates the performance of three multilingual models, i.e., mBERT, XLM-R, and LaBSE in verifying political news claims using the PolitiFact dataset with an integrated multi-evidence approach. The evaluation results reveal notable variations in model performance across multiple metrics, offering insights into their respective strengths, limitations, and suitability for different multi-evidence-based fact verification scenarios.

The findings indicate that under the evaluated PolitiFact multi-evidence setting, LaBSE demonstrates the most balanced performance across sensitivity, precision, and accuracy, making it a reliable choice within the scope of this study. LaBSE achieved the highest accuracy of 86.46% and the best precision of 87.02%, demonstrating its ability to handle semantic similarity and effectively filter irrelevant information. This balance makes LaBSE particularly suitable for real-world deployment scenarios that require high decision reliability, such as policy analysis, content moderation enforcement, and fact verification systems where false positives must be minimized. Conversely, mBERT, despite having the highest sensitivity of 89.44%, exhibited lower precision (81.45%). This trade-off suggests that the model is better suited for applications prioritizing broad detection coverage rather than strict decision accuracy. In practical terms, such characteristics are advantageous for early-stage misinformation monitoring or large-scale content screening systems, where maximizing recall is critical and false positives can be filtered in subsequent verification stages. However, this behavior makes mBERT less suitable for high-stakes applications that require precise and authoritative verification outcomes. Meanwhile, XLM-R, with the lowest sensitivity of 78.81%, demonstrated the best performance stability with the lowest standard deviation (0.4919). This stability indicates that XLM-R may be appropriate for deployment scenarios that require consistent and predictable behavior across diverse inputs, although it may be less effective when handling claims that require complex or extensive evidence aggregation.

The implementation of the integrated multi-evidence approach in this study has proven to be effective in enabling models to analyze the relationship between claims and multiple pieces of evidence using the "[SEP]" token. However, several challenges were identified, including information redundancy, conflicts among evidence, and input length, which can affect the model's efficiency and accuracy. Consequently, further research is required to optimize evidence selection strategies and input handling mechanisms to enhance robustness and scalability for real-world long-text, multi-evidence fact verification tasks.

6. Declarations

6.1. Author Contributions

Conceptualization: K.; Methodology: E.U.; Software: N.A.; Data Curation: N.A.; Investigation: N.A.; Formal Analysis: T.H.; Validation: E.U.; Writing, Original Draft Preparation: N.A.; Writing, Review and Editing: K., E.U., and T.H.; Supervision: K.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The PolitiFact dataset employed in this research is publicly accessible through https://www.politifact.com.

6.3. Funding

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) scheme, funded under Contract No. 126/C3/DT.05.00/PL/2025 dated May 28, 2025, and Decree No. 0419/C3/DT.05.00/2025 dated May 22, 2025. The project entitled "Pengembangan Seleksi Fitur pada Representasi Graf untuk Penyaringan Bukti dalam Verifikasi Fakta Berita Politik Bahasa Indonesia" was headed by Kusrini and implemented through a collaborative partnership 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). The authors sincerely acknowledge the valuable contributions and support of all collaborating institutions and research team members.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Bucciol, "False claims in politics: Evidence from the US," *Research in Economics*, vol. 72, no. 2, pp. 196–210, Jun. 2018, doi: 10.1016/j.rie.2018.04.002.

[2] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, and P. Nakov, "Proppy: Organizing the news based on their propagandistic content," *Inf. Process. Manag*., vol. 56, no. 5, pp. 1849–1864, Sep. 2019, doi: 10.1016/j.ipm.2019.03.005.

[3] N. Agustina, Kusrini, E. Utami, and T. Hidayat, "Systematic Literature Review in the Development of Datasets and Fact Verification Models for Indonesian Language," i*n 2024 7th International Conference of Computer and Informatics Engineering (IC2IE), IEEE,* vol. 2024, no. Sep., pp. 1–9, 2024. doi: 10.1109/IC2IE63342.2024.10748079.

[4] D. Calvo, L. Valera-Ordaz, M. Requena i Mora, and G. Llorca-Abad, "Fact-checking in Spain: Perception and trust," *Catalan Journal of Communication & Cultural Studies*, vol. 14, no. 2, pp. 287–305, Oct. 2022, doi: 10.1386/cjcs_00073_1.

[5] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 2021, no. Jan., pp. 1–15, 2021, doi: 10.1007/s11042-020-10183-2.

[6] M. Larki and E. Manouchehri, "Dispelling Fake News and Infodemic Management about COVID-19 Vaccination: A Literature Review NewsandInfodem," *Journal of Health Literacy*, vol. 7, no. 3, pp. 91–105, 2022, doi: 10.22038/jhl.2022.65215.1289.

[7] C. Chen, W. Chen, J. Zheng, A. Luo, F. Cai, and Y. Zhang, "Input-oriented demonstration learning for hybrid evidence fact verification," *Expert Syst. Appl*., vol. 246, no. Jul., pp. 123191–123191, 2024, Jul. 2024, doi: 10.1016/j.eswa.2024.123191.

[8] 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, no. Aug., pp. 86859–86871, 2023, doi: 10.1109/ACCESS.2023.3304915.

[9] 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,* vol. 2023, no. Oct., pp. 13377–13392, 2023. doi: 10.18653/v1/2023.emnlp-main.826.

[10] Y. Yang, Y. Zhou, Q. Ying, Z. Qian, and X. Zhang, "Search, Examine and Early-Termination: Fake News Detection with Annotation-Free Evidences," in *Frontiers in Artificial Intelligence and Applications,* vol. 392, no. Jan., pp. 1463–1470, 2024. doi: 10.3233/FAIA240649.

[11] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, "Evidence-aware Fake News Detection with Graph Neural Networks," in *Proceedings of the ACM Web Conference 2022, New York, NY, USA: ACM*, vol. 2022, no. Apr., pp. 2501–2510, 2022. doi: 10.1145/3485447.3512122.

[12] N. Vo and K. Lee, "Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Stroudsburg, PA, USA: Association for Computational Linguistics*, vol. 2021, no. Feb., pp. 965–975, 2021. doi: 10.18653/v1/2021.eacl-main.83.

[13] L. Silva and L. Barbosa, "Improving dense retrieval models with LLM augmented data for dataset search," *Knowl. Based. Syst.*, vol. 294, no.1, p. 111740, Jun. 2024, doi: 10.1016/j.knosys.2024.111740.

[14] X. Zhang and W. Gao, "Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method," 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*, vol. 2023, no. Sep., pp. 996–1011, 2023. doi: 10.18653/v1/2023.ijcnlp-main.64.

[15] N. Tiyajamorn, T. Kajiwara, Y. Arase, and M. Onizuka, "Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics,* vol. 2021, no. Nov., pp. 7764–7774, 2021. doi: 10.18653/v1/2021.emnlp-main.612.

[16] 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.

[17] G. Li, Z. Wang, M. Zhao, Y. Song, and L. Lan, "Sentiment Analysis of Political Posts on Hong Kong Local Forums Using Fine-Tuned mBERT," in *Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022, Institute of Electrical and Electronics Engineers Inc.*, vol. 2022, no. Dec., pp. 6763–6765, 2022. doi: 10.1109/BigData55660.2022.10020704.

[18] A. Chauhan, T. Agrawal, and A. Singh, "Advancing Linguistic Frontiers with mBERT Fine-Tuning for Hindi and English Named Entity Recognition Using HiNER and WikiNEuRal Datasets," in *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024, Institute of Electrical and Electronics Engineers Inc.*, vol. 2024, no. Jul., pp. 1–10, 2024. doi: 10.1109/ICCCNT61001.2024.10724051.

[19] N. Rathod, N. Mistry, D. Talati, M. Parikh, A. Kore, and P. Kanani, "Marathi Social Media Opinion Mining using XLM-R," in *Proceedings - International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2022, Institute of Electrical and Electronics Engineers Inc.*, vol. 2022, no. Jul., pp. 730–736, 2022. doi: 10.1109/ICAAIC53929.2022.9793308.

[20] N. Rajapaksha, S. Ahangama, and S. Adikari, "Fine-tuning XLM-R for the Detection of Sinhala Hate Speech Content on Twitter and Youtube," in *ICARC 2023 - 3rd International Conference on Advanced Research in Computing: Digital Transformation for Sustainable Development, Institute of Electrical and Electronics Engineers Inc.*, vol. 2023, no. Feb., pp. 19–23, 2023. doi: 10.1109/ICARC57651.2023.10145745.

[21] G. Mehak, I. Muneer, and R. M. A. Nawab, "Urdu Text Reuse Detection at Phrasal level using Sentence Transformer-based approach," *Expert Syst. Appl.,* vol. 234, no.1, p. 121063, Dec. 2023, doi: 10.1016/j.eswa.2023.121063.

[22] H. Ma ., "EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification," in *Findings of the Association for Computational Linguistics ACL 2024, Stroudsburg, PA, USA: Association for Computational Linguistics*, vol. 2024, no. Oct., pp. 9340–9353, 2024. doi: 10.18653/v1/2024.findings-acl.556.

[23] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics*, vol. 2019, no. Jul., pp. 4996–5001, 2019. doi: 10.18653/v1/P19-1493.

[24] A. Conneau ., "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*, vol. 2020, no. Jul., pp. 8440–8451, 2020. doi: 10.18653/v1/2020.acl-main.747.

[25] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 2022, no. Jul., pp. 878–891, 2022, doi: 10.18653/v1/2022.acl-long.62.

[26] S. Shukla, H. Dutta, and P. Bhattacharyya, "Recon, Answer, Verify: Agents in Search of Truth," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Stroudsbur*g, PA, USA: Association for Computational Linguistics, vol. 2025, no. Jul., pp. 2429–2448, 2025. doi: 10.18653/v1/2025.emnlp-industry.167.

[27] M. Soprano ., "The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale," *Inf. Process. Manag.*, vol. 58, no. 6, p. 102710, Nov. 2021, doi: 10.1016/j.ipm.2021.102710.

[28] 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.*, vol. 2021, no. Apr., pp. 241–246, 2021. doi: 10.1109/ICAIIC51459.2021.9415192.

[29] Y. Li ., "Deep learning-based platform performs high detection sensitivity of intracranial aneurysms in 3D brain TOF-MRA: An external clinical validation study," *Int. J. Med. Inform.*, vol. 188, no. Aug., pp. 105487–105487, doi: 10.1016/j.ijmedinf.2024.105487.

[30] S. Jamshidi ., "Effective text classification using BERT, MTM LSTM, and DT," *Data Knowl. Eng.*, vol. 151, no. May, pp. 102306–102306, 2024, doi: 10.1016/j.datak.2024.102306.

[31] 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, no.1, pp. 902–907, 2024, doi: 10.1016/j.procs.2024.06.110.