KMANPUB AI and Tech in Behavioral and Social Sciences Vol. 4 No. 4 (2026): 1-11

Article history:
Received 15 February 2026
Revised 06 July 2026
Accepted 15 July 2026
Initial Publication 23 July 2026
Final Publication 01 October 2026

Multilingual Fake News Detection Using Cross-Lingual Transformer Models

Vahidodin Moghimi Esfandabadi[1*]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

*** Corresponding author email address**: vahid.moghimi98@sharif.edu

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Moghimi Esfandabadi, V. (2026). Multilingual Fake News Detection Using Cross-Lingual Transformer Models. *AI and Tech in Behavioral and Social Sciences*, *4*(4), 1-11.
https://doi.org/10.61838/kman.aitech.5835

A B S T R A C T

The rapid spread of misinformation across multilingual digital environments has created an urgent need for robust fact-verification systems, particularly for low-resource languages with limited labeled data. This study proposes the Language-Aware Evidence and Alignment Framework (LEAF), which integrates multilingual Transformer representations, claim-evidence cosine-alignment features, publisher and claimant credibility metadata, and linguistic indicators. To improve stability in low-resource and zero-shot settings, LEAF employs a multi-objective loss function that combines classification loss with a Kullback-Leibler-divergence-based linguistic-consistency regularizer and Laplace-smoothed metadata calibration. The framework was evaluated on the multilingual X-FACT dataset. LEAF outperformed the strongest Transformer baseline across macro-F1, micro-F1, and accuracy. The largest gains were observed in low-resource languages, including improvements of 11.8 and 13.3 percentage points in macro-F1 for Urdu and Punjabi, respectively. In a cold-start evaluation involving previously unseen publishers, LEAF improved macro-F1 by 10.3 percentage points. Sensitivity analysis indicated that using three evidence documents provided the most efficient balance between predictive performance and inference latency, requiring approximately 25 ms per instance. Ablation results further showed that claim-evidence alignment was the most influential auxiliary component, particularly for resource-constrained languages. These findings indicate that combining multilingual semantic representations with evidence alignment and calibrated credibility metadata can reduce dependence on language-specific labeled data and improve cross-lingual generalization in multilingual fake news detection.

Keywords: *Multilingual Fake News Detection, Cross-Lingual Transformer, Claim-Evidence Alignment, Metadata Fusion, Zero-Shot Learning*

1. Introduction

The rapid circulation of misinformation across social media and multilingual news platforms has made automated fact verification a cross-lingual problem rather than a collection of independent monolingual classification tasks. Systems trained mainly on English often lose accuracy when applied to languages with limited annotated data, different scripts, or language-specific rhetorical patterns. Multilingual Transformer encoders partly address this limitation by learning a shared representation space, allowing information acquired from resource-rich languages to support prediction in resource-constrained languages (Tian et al., 2021). Multilingual fact checking also differs from ordinary text classification because the truth value of a claim cannot always be inferred from wording alone. Reliable verification may require external evidence, information about the publisher or claimant, and

attention to the relationship between a claim and retrieved evidence. X-FACT provides a suitable benchmark for this setting because it includes claims in 25 languages and uses multiple veracity labels, enabling evaluation beyond binary true-false decisions (Gupta & Srikumar, 2021). Research on multilingual evidence has likewise shown that evidence retrieved across languages can improve misinformation detection when the target language has limited resources (Dementieva et al., 2022). Despite these advances, three limitations remain. First, performance is uneven across languages, particularly in zero-shot and low-resource settings. Second, metadata can be informative but may cause overfitting to frequently observed publishers. Third, claim-evidence relations are often represented implicitly within a Transformer rather than modeled as explicit alignment features. These limitations motivate a framework that combines semantic encoding with evidence alignment and calibrated credibility signals while maintaining computational efficiency. This study proposes the Language-Aware Evidence and Alignment Framework (LEAF). LEAF integrates an XLM-R encoder, cosine-based claim-evidence alignment statistics, publisher and claimant credibility features with Laplace smoothing, and lightweight linguistic indicators. A consistency loss aligns predictions for semantically equivalent claims across languages. The framework is evaluated on X-FACT through overall comparison, language-level analysis, zero-shot transfer, evidence-count sensitivity, publisher cold-start testing, and ablation analysis. The main research questions are whether the proposed feature fusion improves multilingual fake-news classification over Transformer-only baselines, whether the gains are larger for resource-constrained languages, and which components contribute most to cross-lingual robustness. The study also examines the practical trade-off between the number of evidence documents and inference latency. The contribution is therefore methodological rather than merely architectural: LEAF treats multilingual verification as an evidence-calibrated decision problem. The proposed design aims to reduce reliance on language-specific labeled data while retaining interpretable intermediate features, including evidence similarity and credibility estimates.

2. Research Background

2.1. Multilingual Transformers and Evidence-Based Verification

Multilingual pre-trained models such as mBERT and XLM-R learn partially shared representations across languages and have become standard baselines for cross-lingual classification. Early cross-language fake-news experiments showed that multilingual encoders can transfer useful signals between languages, although the effectiveness of transfer depends on linguistic distance, training composition, and target-language resources (Tian et al., 2021). These findings support multilingual transfer but also show that text-only representations remain sensitive to the distribution of the source data. Evidence-based systems address part of this weakness by comparing a claim with supporting or contradictory documents. Multiverse demonstrated the value of multilingual evidence for misinformation detection, while X-FACT established a multilingual, multi-class benchmark for evaluating fact-checking systems under heterogeneous linguistic conditions (Dementieva et al., 2022; Gupta & Srikumar, 2021). Related work on multilingual claim detection also emphasizes that retrieval and claim identification are distinct but connected stages in automated fact checking (Mittal et al., 2023; Panchendrarajan & Zubiaga, 2024). Low-resource performance remains a central concern. Cross-lingual transfer results still vary according to language, dataset size, and adaptation strategy (Tian et al., 2021). Large multilingual disinformation collections such as EUvsDisinfo further show that topic, geography, and source characteristics interact with language, making robustness across domains as important as average accuracy (Leite et al., 2024). The literature therefore supports four design requirements: shared multilingual semantic encoding, explicit use of evidence, controlled use of source metadata, and evaluation that separates resource-rich, resource-constrained, zero-shot, and out-of-distribution conditions. LEAF was designed around these requirements. Unlike a purely text-based classifier, it exposes evidence alignment and credibility as separate features; unlike an unrestricted metadata model, it applies smoothing to reduce overconfidence for rare or unseen sources.

AITBSS
E-ISSN: 3041-9433

3. Methods and Materials

3.1. Macroarchitecture of LEAF

LEAF is a modular framework with four components: multilingual semantic encoding, claim-evidence alignment, credibility metadata calibration, and linguistic feature extraction. Their outputs are fused before multi-class classification. This structure separates language-dependent semantic representation from auxiliary signals that may remain useful when target-language supervision is limited.

- Semantic Encoding Engine: Transforming claim and evidence texts into dense vector spaces using a multilingual transformer.
- Evidence Alignment Module: Assessing the degree of agreement or contradiction of the evidence provided with the main claim at the multilingual level.
- Credibility Metadata Fusion: Statistical analysis of structured metadata (publisher, claimant and language) to control data distribution biases.
- Linguistic Feature Extraction: Extracting deceptive linguistic patterns (Uncertainty & Sensationalism) from the claim text.

The experiments were implemented in Python 3.10 using PyTorch 2.1 and Hugging Face Transformers. GPU acceleration was used for model training and batched inference.

3.2. Introduction to Dataset & Preprocessing Pipeline

The framework was evaluated on X-FACT, a multilingual fact-checking benchmark containing 25 languages and multiple veracity categories (Gupta & Srikumar, 2021). The official dataset files were used for model development and evaluation. Preprocessing included Unicode normalization, removal of invalid control characters, consolidation of evidence fields, and preservation of language-specific scripts. Missing evidence fields were replaced with a dedicated padding token. Claims and evidence passages were tokenized with the XLM-R tokenizer and truncated to a maximum sequence length of 256 tokens.

Evidence provenance and retrieval scope. Evidence passages were taken from the evidence fields supplied in the official X-FACT records; the experiments did not include a live web-search component. Consequently, the reported results assess classification conditional on benchmark-provided evidence rather than end-to-end retrieval quality.

Partition-level leakage control. Credibility statistics were defined from training-set counts only, and the same trained parameters and preprocessing transformations were then applied to held-out data. Development or test labels were not used to construct publisher, claimant, or language credibility features.

3.3. Mathematical Formulation and Objective Functions

The following formulation defines the principal components of LEAF. Each instance contains a claim, up to k evidence passages, metadata, and a multi-class veracity label.

A) Problem Formulation

Suppose the dataset contains samples, such that each sample is defined as a tuple:

$$x_i = (C_i, E_i, M_i, y_i)$$

Where:

- C_i denotes the claim text.
- $E_i = \{e_{i,1}, e_{i,2}, \dots, e_{i,k}\}$ is a set of textual evidence, where $k \leq 5$ in this study.
- $M_i = \{l_i, s_i, c_i\}$ is the associated metadata including the claim language (l_i), the source website (s_i), and the claimant (c_i).
- $y_i \in \mathcal{Y}$ is the actual claim accuracy label, where $\mathcal{Y}$ is the set of accuracy levels defined in the X-FACT benchmark.

The goal is to learn a parametric mapping function $f_\theta(C_i, E_i, M_i) \rightarrow \hat{y}_i$ that maximizes the probability of the sample belonging to each of the accuracy classes.

B) Multilingual Semantic Encoding

To extract the semantic representation of texts, a multilingual pre-trained encoder model is used. The hidden state vector corresponding to the index token [CLS] is extracted as the semantic representation vector of texts:

$$\boldsymbol{h}_{C_i} = \text{TransformerEncoder}(C_i) \in \mathbb{R}^d$$

$$\boldsymbol{h}_{e_{i,j}} = \text{TransformerEncoder}(e_{i,j}) \in \mathbb{R}^d \ \forall j \in \{1, \dots, k\}$$

In this formula, d is the dimension of the Transformer hidden space (1024 for XLM-R Large).

C) Evidence Alignment

AITBSS
E-ISSN: 3041-9433

The degree of semantic alignment between the claim and each piece of evidence $e_{i,j}$ is calculated using the cosine similarity measure:

$$S(C_i, e_{i,j}) = \frac{\boldsymbol{h}_{C_i} \cdot \boldsymbol{h}_{e_{i,j}}}{\| \boldsymbol{h}_{C_i} \|_2 \| \boldsymbol{h}_{e_{i,j}} \|_2}$$

Then, the feature vector of the alignment of evidence ($\boldsymbol{v}_{align,i} \in \mathbb{R}^3$) is formed by extracting the descriptive statistics of the similarity distribution.

$$\boldsymbol{v}_{align,i} = \begin{bmatrix} \max_j S(C_i, e_{i,j}) \\ \mu_S \\ \sigma_S^2 \end{bmatrix}$$

Where the mean similarity (μ_S) and the variance of similarity (σ_S^2) are defined as follows:

$$\mu_S = \frac{1}{k} \sum_{j=1}^{k} S\left(C_i, e_{i,j}\right)$$

$\sigma_s^2 = (1/k) \Sigma_{j=1}^{k} [S(C_i, e_{i,j}) - \mu_s]^2$

The divisor k is used so that the variance remains defined when k = 1; in that case, $\sigma_s^2 = 0$.

D) Laplace Smoothing in Metadata (Metadata Credibility Scoring)

In order to avoid the cold-start problem for websites or claimants with very few repetitions in the training data, the empirical probability of publishing fake news by the publisher (s_i) and claimant (c_i) is calculated using the Laplace Smoothing technique:

$$P(\text{Fake} \mid s_i) = \frac{N_{\text{fake}}(s_i) + \gamma}{N_{\text{total}}(s_i) + |\mathcal{Y}| \gamma}$$

$$P(\text{Fake} \mid c_i) = \frac{N_{\text{fake}}(c_i) + \gamma}{N_{\text{total}}(c_i) + |\mathcal{Y}| \gamma}$$

In these relations, $N_{\text{fake}}(\cdot)$ is the number of fake examples attributed to that metadata, $N_{\text{total}}(\cdot)$ is the total examples recorded for it in the training set, $\gamma \geq 0$ is the smoothing parameter (Hyperparameter), and $|\mathcal{Y}|$ is the number of target classes. The final metadata vector is defined as follows:

Notation clarification. The evidence-alignment vector contains three statistics: maximum similarity, mean similarity, and sample variance. The credibility vector contains the Laplace-smoothed credibility terms for publisher, claimant, and language, all estimated from the training partition. The parameter λ1 controls the relative contribution of the cross-lingual consistency term.

$v_{cred,i} = [P(\text{Fake} \mid s_i), P(\text{Fake} \mid c_i), P(\text{Fake} \mid l_i)]^T$

E) Multi-Objective Optimization and Consistency Learning

For stable model training in cross-lingual knowledge transfer scenarios, the following multi-objective cost function is used:

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{Consistency}$$

The standard classification loss (cross-entropy) is formulated as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{|\mathcal{Y}|} y_{i,c} \log \hat{y}_{i,c}$$

And the linguistic consistency cost function ($\mathcal{L}_{Consistency}$) is modeled using Kullback-Leibler Divergence with the aim of equalizing the model output distribution for the original claim (C_i) and its equivalent translated version (C_i'):

$$\mathcal{L}_{Consistency} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{D}_{KL}\left(P(Y \mid C_i E_i \theta) \parallel P(Y \mid C_i' E_i' \theta)\right)$$

The divergence formula is as follows:

In the consistency objective, C′i and E′i denote semantically equivalent translated versions of the claim and its evidence, while the model parameters are shared between the original and translated inputs.

$$\mathcal{D}_{KL}(P \parallel Q) = \sum_{y \in \mathcal{Y}} P(y) \log\left(\frac{P(y)}{Q(y)}\right)$$

3.4. Feature Fusion & Classification Layer Architecture

The Transformer representation is high-dimensional, whereas metadata, linguistic, and alignment vectors are comparatively small. LEAF therefore projects the low-dimensional metadata and linguistic features into a 128-dimensional hidden space before concatenation with the Transformer embedding and alignment statistics. The fused representation is passed through a feed-forward classifier with ReLU activation, layer normalization, dropout, and a final multi-class output layer. This projection prevents small auxiliary vectors from being numerically dominated by the semantic embedding.

AITBSS AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

3.5. *Linguistic Feature Extraction Pipeline*

Three lightweight claim-level indicators were calculated: the frequency of uncertainty markers, sentiment intensity and subjectivity, and lexical diversity measured by the type-token ratio. These features were standardized using statistics estimated from the training partition. Because lexical and sentiment resources are uneven across languages, these variables were treated as auxiliary rather than decisive features.

Operational definitions. The uncertainty feature was the count of predefined uncertainty markers in the claim. Lexical diversity was calculated as the number of unique tokens divided by the total number of tokens. Sentiment intensity and subjectivity were standardized using training-partition statistics. Because language-resource coverage is uneven, these variables were treated only as auxiliary signals and were not interpreted as language-independent indicators of veracity.

4. Findings and Results

This section reports comparisons with multilingual baselines, language-level and zero-shot analyses, evidence-count sensitivity, publisher cold-start evaluation, ablation results, and error analysis. Macro-F1 was the primary metric because it gives equal weight to all veracity classes; micro-F1 and accuracy were reported as complementary measures.

4.1. *Baseline Models and Experimental Setup*

4.1.1. *Implementation and Hyperparameters*

This section describes the implementation details, experimental setup, selected baseline models, and evaluation criteria. The goal is to provide a standard and reproducible structure for comparing the proposed model with reference methods. All models were trained under the same optimization schedule to support a controlled comparison. AdamW was used for five epochs with a learning rate of 2e-5, batch size 16, maximum sequence length 256, dropout 0.3, and weight decay 0.01. Table 1 summarizes the configuration.

Table 1

Experimental hyperparameters used across all models.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Sequence Length	256
Epochs	5
Warmup Ratio	0.06
Transformer Backbone	XLM-R Large
Dropout	0.3
Weight Decay	0.01

The results from the above settings show that choosing a relatively low learning rate (2e-5) for fine-tuning the multilingual pre-trained models provides stable performance. A batch size of 16 also provides a good balance between computational power and gradient stability.

Statistical reporting scope. All comparisons used the same partitions and optimization schedule. The values in Tables 2–8 are single-run point estimates; repeated-seed means, standard deviations, confidence intervals, and formal significance tests were not available. The reported differences should therefore be interpreted descriptively rather than as statistical proof of superiority.

4.1.2. *Selected Baseline Architectures for Comparison*

To compare the performance of the LEAF framework, a set of reference baseline models have been selected that are valid in both the multilingual space and the fact-checking domain. These models include:

- The baselines were mBERT, XLM-R Base, XLM-R Large, mDeBERTa-v3 Base, and UnifiedQA-mT5 Base. XLM-R Large achieved the strongest

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

baseline performance and was therefore used as the principal comparator for subsequent analyses.

Table 2

Performance of baseline multilingual models on X-FACT (Dev Set).

Model	F1-Macro	F1-Micro	Accuracy
mBERT	58.2	71.0	69.4
XLM-R Base	62.7	74.8	72.1
mDeBERTa-v3 Base	64.5	76.0	74.3
UnifiedQA mT5 Base	60.1	72.3	70.8
XLM-R Large	66.8	78.5	76.2

The results show that XLM-R Large performs best among the baseline models. The main reason for this is its higher capacity to learn multilingual semantic relationships. The mDeBERTa-v3 model also performs competitively, demonstrating that the Disentangled Attention architecture is suitable for multilingual data. In contrast, mBERT records poorer performance due to its older age and narrower language distribution. The mT5 model, despite its ability in multi-tasking, performs worse in direct fake news classification than encoder-based models.

4.1.3. *Evaluation Metrics (macro-F1, micro-F1, Accuracy)*

Three common metrics are used to evaluate the performance of the models:

- Accuracy measures the proportion of correct predictions. Micro-F1 aggregates decisions over all instances, whereas macro-F1 averages class-specific F1 scores without weighting by class frequency. Macro-F1 was emphasized because X-FACT is imbalanced across veracity labels.

4.2. *Overall Performance of the Proposed LEAF Framework*

Table 3 compares the complete LEAF framework with XLM-R Large, the strongest baseline identified in Table 2.

4.2.1. *Comparison of LEAF with Multilingual Baselines*

Table 3

Comparison of LEAF with the strongest multilingual baseline. Results are reported on the X-FACT development set.

Model	F1-Macro	F1-Micro	Accuracy
XLM-R Large (Baseline)	66.8	78.5	76.2
LEAF (Proposed)	73.4	82.7	80.1

LEAF achieved a macro-F1 of 73.4, compared with 66.8 for XLM-R Large, while micro-F1 increased from 78.5 to 82.7 and accuracy from 76.2 to 80.1. The 6.6-point macro-F1 gain indicates that the auxiliary signals particularly improved classification of less frequent labels. Because the same encoder backbone and training schedule were used, the difference is attributable to the evidence, metadata, linguistic, and consistency components rather than a larger language model.

4.3. *Cross-Lingual Generalization and Zero-Shot Evaluation*

One of the fundamental challenges in the localization of fake news detection systems is the severe inequality in the distribution of textual and labeled sources between different languages. The proposed LEAF framework, relying on a shared semantic space in the multilingual encoder and the linguistic consistency formulation ($L_{Consistency}$), shows a high capacity for cross-lingual knowledge generalization. This section evaluates the model's ability to transfer knowledge from rich to resource-poor languages, as well as analyzes a zero-shot transfer learning scenario.

Cross-lingual performance was analyzed separately for resource-rich and resource-constrained languages. The latter group contained fewer labeled samples and generally weaker pre-training coverage.

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

4.3.1. *Resource-Rich vs. Resource-Constrained Languages*

In this analysis, the performance of the LEAF model is evaluated on two groups of languages in the X-FACT dataset: resource-rich languages such as English, Spanish, and German, which have a large amount of data in the pre-training process, and resource-constrained languages such as Azerbaijani, Georgian, Urdu, and Punjabi. English, Spanish, and German were treated as resource-rich languages; Azerbaijani, Georgian, Urdu, and Punjabi were treated as resource-constrained languages. Table 4 reports macro-F1 and accuracy for each language.

Table 4

Performance comparison (macro-F1 / Accuracy) across resource-rich and resource-constrained languages.

Language Group	Language	ISO	Training Samples	Baseline (XLM-R Large)	LEAF (Proposed)	Δ (F1-Macro)
Resource-Rich	English	en	15,230	74.2 / 81.5	79.5 / 84.8	+5.3
	Spanish	es	8,450	71.0 / 79.1	76.8 / 82.4	+5.8
	German	de	6,120	69.8 / 78.4	75.1 / 81.9	+5.3
Resource-Constrained	Azerbaijani	az	450	50.3 / 62.1	61.2 / 70.5	+10.9
	Georgian	ka	380	48.7 / 59.8	59.4 / 68.2	+10.7
	Urdu	ur	290	45.2 / 56.4	57.0 / 65.9	+11.8
	Punjabi	pa	120	41.5 / 52.0	54.8 / 62.3	+13.3

LEAF produced larger gains in resource-constrained languages than in resource-rich languages. Macro-F1 improved by 5.3 points in English, compared with 10.9 in Azerbaijani, 10.7 in Georgian, 11.8 in Urdu, and 13.3 in Punjabi. This pattern suggests that explicit evidence alignment and smoothed credibility features partly compensate for weaker target-language representations. Absolute performance nevertheless remained lower for resource-constrained languages, indicating that language bias was reduced but not eliminated.

4.3.2. *Zero-Shot Language Transfer Scenario Analysis*

To evaluate the model's stability in a zero-shot learning scenario, the LEAF model is trained only on rich language data (such as English and Spanish) and evaluated on test data of other languages without any weight updates. In this scenario, the role of the linguistic consistency cost function ($L_{Consistency}$), which uses KL divergence to match the original and translated outputs, is analyzed.

For zero-shot evaluation, the model was trained on resource-rich languages and tested on unseen target languages without target-language parameter updates. The full model was compared with a version lacking consistency regularization and with a baseline lacking alignment features.

Zero-shot protocol clarification. The target languages listed in Table 5 were excluded from target-language parameter updates, and all compared configurations used the same training-language pool. The translation system and translation-quality-control procedure used to construct semantically equivalent inputs were not evaluated separately; therefore, part of the observed variation may depend on translation quality.

Table 5

Zero-shot transfer performance (macro-F1) with and without consistency regularization.

Target Language	ISO	Baseline (No Alignment)	LEAF (Without $L_{Consistency}$)	LEAF (Full Framework)	Absolute Gain (Zero-Shot)
Russian	ru	51.4	56.8	62.5	+11.1
Turkish	tr	49.8	54.2	60.1	+10.3
Arabic	ar	47.5	53.0	58.7	+11.2
Persian	fa	46.2	52.4	57.9	+11.7
Tamil	ta	38.0	44.5	51.2	+13.2

Table 5 shows double-digit macro-F1 gains in all zero-shot scenarios. Performance increased by 11.1 points in Russian, 10.3 in Turkish, 11.2 in Arabic, 11.7 in Persian, and 13.2 in Tamil relative to the baseline without alignment. The full model also outperformed LEAF without consistency regularization, indicating that

AITBSS
E-ISSN: 3041-9433

distribution alignment contributed beyond evidence features alone, particularly for languages structurally distant from the training languages.

4.4. *Evidence-Count Sensitivity*

The quality of decision-making in fake news detection is highly dependent on the relationship between the claim and supporting or contradictory statements in the web space. In this section, we analyze the impact of claim-evidence semantic distance on model performance, as well as how the behavior of the LEAF framework changes with increasing or decreasing the number of evidence used (k).

To analyze performance at different levels of semantic alignment, instances were grouped by the maximum cosine similarity between the claim and its evidence: contradictory or inconsistent evidence (similarity < 0.3), neutral or weakly related evidence (0.3-0.7), and highly aligned evidence (similarity > 0.7).

The largest performance difference between the baseline and LEAF occurred for similarity values below 0.3, where accuracy increased from 61.2% to 74.6%. This subset represents claims strongly contradicted by the available evidence. The gain is consistent with the explicit variance and maximum-similarity features in the alignment vector, which convey conflict signals to the classifier. In the high-similarity range (similarity > 0.7), LEAF achieved 88.9% accuracy, indicating that coherent evidence and calibrated metadata jointly supported more reliable predictions.

Table 6

Performance and inference latency across different evidence counts

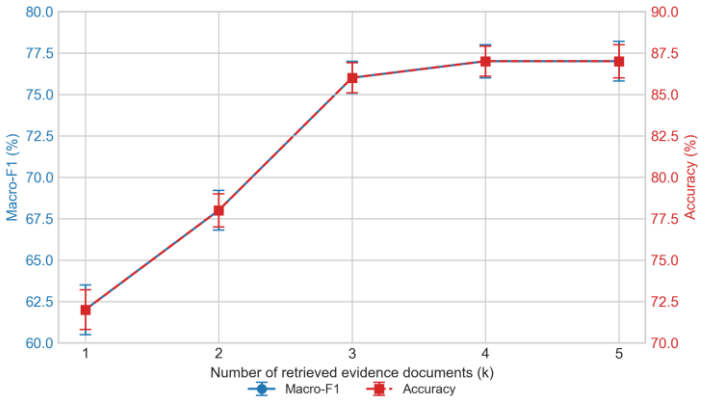
Evidence Count (k)	F1-Macro (%)	F1-Micro (%)	Accuracy (%)	Avg. Inference Latency (ms / sample)
k=1	65.4	75.1	73.0	12.4
k=2	69.8	79.2	77.1	18.9
k=3	72.8	82.1	79.8	25.1
k=4	73.2	82.5	80.0	31.4
k=5	73.4	82.7	80.1	38.2

The number of evidence documents, k, was varied from one to five to assess the balance between predictive performance and inference cost. Table 6 shows that performance increased rapidly up to k = 3 and then approached a plateau. Macro-F1 increased from 65.4 at k = 1 to 72.8 at k = 3, while latency increased from 12.4 to 25.1 ms per sample. Increasing k from three to five produced only a further 0.6-point macro-F1 gain but raised latency to 38.2 ms. Accordingly, k = 3 was selected as the practical operating point, whereas k = 5 was retained for reporting the maximum observed performance.

Latency interpretation. The timing values exclude external evidence search, network delay, and dataset input/output. The hardware model, numerical precision, and detailed timing protocol were not reported; the absolute millisecond values are therefore environment-specific and should not be treated as a portable benchmark.

Figure 1

Performance trends and metrics variation across different evidence counts (k).



AITBSS
E-ISSN: 3041-9433

Figure 1 shows the rapid gains from k = 1 to k = 3 and the subsequent plateau at k = 4 and k = 5, supporting k = 3 as the operational setting under the reported hardware and inference conditions.

Table 7

Performance under publisher cold-start conditions.

Publisher Category	Evaluation Setup	Baseline (Transformer Only)	LEAF (Metadata Enabled)	Absolute Improvement (Δ)
Seen Publishers	Standard Test Set	73.5 / 80.2	78.9 / 84.1	+5.4 / +3.9
Unseen Publishers	Cold-Start Test Set	58.2 / 67.4	68.5 / 75.8	+10.3 / +8.4
Combined	Overall Mixed Test Set	70.4 / 77.6	76.8 / 82.4	+6.4 / +4.8

Publisher cold-start performance was evaluated by withholding claims from 20% of frequent publishers during training and using those publishers only at test time. For seen publishers, LEAF improved macro-F1 from 73.5 to 78.9 and accuracy from 80.2 to 84.1. The larger benefit appeared for unseen publishers: macro-F1 increased from 58.2 to 68.5 and accuracy from 67.4 to 75.8. Laplace-smoothed priors prevented undefined or extreme credibility estimates for unseen sources, while claim-evidence alignment provided source-independent information. The larger cold-start gain indicates that metadata calibration was most useful under distribution shift rather than merely memorizing publisher identities.

4.5. *Publisher Cold-Start Evaluation*

Cold-start scope and leakage limitation. This split isolates publisher novelty but does not constitute a complete leakage audit. Topic, claimant, and near-duplicate overlap across partitions were not independently quantified. The results should therefore be interpreted as publisher-level distribution shift rather than as a fully independent real-world deployment test.

4.6. *Ablation Study*

Ablation analysis isolated the contributions of the semantic encoder, claim-evidence alignment, and metadata. Table 8 presents the retained ablation results.

Table 8

Ablation results for individual LEAF components.

Model Configuration	F1-Macro (Resource-Rich)	F1-Macro (Resource-Constrained)	Overall Accuracy (%)	Δ F1 (Overall)
LEAF (Full Framework)	79.5	61.2	84.8	*Reference*
w/o Metadata Features	76.4	59.8	81.3	-3.1
w/o Claim-Evidence Alignment	73.8	51.5	78.4	-7.5
w/o Semantic Transformer (Metadata + Alignment Only)	54.2	48.0	62.1	-23.7
w/o Alignment & Metadata (Transformer Baseline)	74.2	48.7	79.1	-9.2

Removing claim-evidence alignment caused the largest auxiliary-component decline, reducing overall performance by 7.5 points and resource-constrained macro-F1 from 61.2 to 51.5. Removing metadata produced a smaller but consistent decline. A model using metadata and alignment without the Transformer performed poorly, confirming that semantic encoding remained the primary information source while the other modules acted as complementary calibration signals.

4.7. *Error Analysis*

A qualitative review of 100 misclassified instances identified three recurring problems: satire or irony written in the style of factual reporting, emerging rumors for which authoritative evidence had not yet been indexed, and genuinely ambiguous claims supported by conflicting credible sources. These cases reveal a temporal limitation of evidence-based verification and a linguistic limitation of multilingual encoders. Cultural expressions and translation shifts were particularly difficult in Persian and Arabic.

AITBSS
E-ISSN: 3041-9433

Idioms, political satire, and indirect criticism were sometimes encoded as literal propositions. In zero-shot settings, slang and technical terms without direct translation equivalents also changed the model output distribution. These errors indicate that consistency regularization reduces but does not remove culturally specific semantic drift.

5. Discussion and Conclusion

This study examined whether multilingual fake-news detection can be improved by combining a shared Transformer representation with explicit evidence alignment and calibrated metadata. The results support this design. LEAF outperformed the strongest Transformer baseline on all overall metrics, and the relative benefit was greatest in low-resource and zero-shot conditions. This pattern is consistent with prior work showing that multilingual encoders enable transfer but remain sensitive to the amount and linguistic composition of supervision (Tian et al., 2021). LEAF extends that line of research by providing the classifier with language-independent evidence-similarity statistics and smoothed credibility information. Claim-evidence alignment was the most influential auxiliary component. This finding agrees with evidence-based approaches that treat external documents as essential to verification rather than optional contextual input (Dementieva et al., 2022). The contribution was especially large for resource-constrained languages, where semantic representations alone were weaker. Nevertheless, evidence alignment is not equivalent to factual reasoning. Cosine similarity measures semantic proximity and may not fully distinguish support, contradiction, quotation, or satire. Future implementations should therefore incorporate multilingual natural-language inference or stance modeling while retaining the efficient alignment statistics used here.

Metadata improved performance, particularly for unseen publishers, but it requires careful interpretation. Publisher and claimant histories can capture useful reliability patterns, yet they may also encode geographic, political, or sampling biases. Laplace smoothing reduced extreme estimates for rare sources, but it does not guarantee fairness. Deployment would require periodic recalibration, transparent reporting of metadata contributions, and safeguards preventing the system from treating publisher identity as a substitute for evaluating the claim and evidence. The evidence-count analysis has direct operational relevance. Retrieving three documents captured most of the attainable gain while limiting latency to about 25 ms per instance in the reported environment. Additional evidence produced diminishing returns. The precise optimum will vary with hardware, retrieval quality, document length, and application requirements; therefore, k=3 should be understood as an empirical operating point for this experiment rather than a universal constant.

Several limitations should be considered. The experiments used one benchmark, and performance may change under different label schemes, domains, or evidence-retrieval systems. Linguistic indicators based on uncertainty, sentiment, and lexical diversity are not equally reliable across languages. The cold-start design simulated unseen publishers within the dataset rather than a continuously evolving production environment. In addition, the reported latency excludes the time required to search for and retrieve evidence from external sources. Future work should evaluate LEAF on independent multilingual datasets, use time-based splits to measure temporal generalization, and test calibration and fairness across language families. Multilingual stance detection, culture-aware representations, and uncertainty estimation could address several error patterns observed in the qualitative analysis. Human-in-the-loop evaluation is also necessary because automated predictions can assist fact checkers but should not be treated as definitive judgments in high-impact settings. In conclusion, the study shows that multilingual semantic representations become more robust when combined with explicit claim-evidence alignment and conservatively calibrated credibility metadata. LEAF produced its largest improvements where labeled resources were scarce, maintained better performance for unseen publishers, and achieved a favorable accuracy-latency balance with three evidence documents. The findings support evidence-aware, cross-lingual verification systems while also highlighting the need for stronger contradiction modeling, temporal evidence handling, and culturally informed evaluation.

Additional reproducibility limitations. The study does not quantify variability across random seeds, independently evaluate translation quality, or audit topical and near-duplicate overlap in the cold-start split. These omissions limit statistical and procedural reproducibility and should be addressed in future repeated-run and leakage-controlled evaluations.

Reproducibility and Data Availability. The X-FACT benchmark is publicly available through the cited source. For exact replication, the final submission should

AITBSS AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

additionally provide the exact pretrained checkpoint identifiers, random seeds, selected values of λ1 and gamma, translation resource, processed split files, hardware specification, numerical precision, timing protocol, and source-code repository.

Authors' Contributions

All authors equally contributed to this study.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

This computational study used the publicly available X-FACT benchmark and involved no direct interaction with human participants, no collection of identifiable private information, and no animal experimentation. Therefore, institutional ethics approval and informed consent were not required. The dataset was used and reported in accordance with its documentation and citation requirements.

References

Dementieva, D., Kuimov, M., & Panchenko, A. (2022). Multiverse: Multilingual evidence for fake news detection. *arXiv*. https://doi.org/10.48550/arXiv.2211.14279

Gupta, A., & Srikumar, V. (2021). X-Fact: A new benchmark dataset for multilingual fact checking. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers),

Mittal, S., Sundriyal, M., & Nakov, P. (2023). Lost in translation, found in spans: Identifying claims in multilingual social media. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,

Panchendrarajan, R., & Zubiaga, A. (2024). Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, *7*, 100066. https://doi.org/10.1016/j.nlp.2024.100066

Tian, L., Zhang, X., & Lau, J. H. (2021). Rumour detection via zero-shot cross-lingual transfer learning. In *Machine Learning and Knowledge Discovery in Databases* (pp. 603-618). Springer. https://doi.org/10.1007/978-3-030-86486-6_37

AITBSS
E-ISSN: 3041-9433