

EmoFuse: Emotion-Aware Fusion and Adaptive Expert Routing for Multi-Domain Arabic Fake News Detection

Fatiha Charef¹, Abdelhafid Zitouni², Mahieddine Djoudi³, Hichem Rahab⁴

¹LIRE Laboratory, University of Constantine 2, Algeria.

²LIRE Laboratory, University of Constantine 2, Algeria.

³TechNE Laboratory, University of Poitiers, France.

⁴ICOSI Laboratory, University of Khenchela, Algeria.

Faculté des Nouvelles Technologies de L'information et de la communication, université abdel-hamid mehri constantine 2

Email:1 fatiha.charef@univ-constantine2.dz, 2abdelhafid.zitouni@univ-constantine2.dz, 3 ma-hieddine.djoudi@univ-poitiers.fr

4rahab.hichem@univ-khenchela.dz

Received : 03/02/2026 ; Accepted : 27/05/2026

Abstract: Arabic fake-news detection must account for linguistic variation and differences between content domains. We propose EmoFuse, a MARBERT-based framework that combines two independently fine-tuned veracity experts with an emotion-specialized contextual representation and a domain-supervised routing network. Each expert fuses a 128-dimensional veracity vector with a 768-dimensional emotion representation, while the gate assigns input-dependent weights to expert predictions. Adaptive routing improves Macro-F1 by an average of 4.78 percentage points over uniform expert averaging. The primary system achieves a macro-F1 of 0.8735 on the historical official test set. The results support adaptive routing across the two observed dataset-defined domains, but they do not provide consistent evidence that emotion specialization adds complementary information beyond veracity-orientated or generic contextual representations.

Keywords: Arabic fake-news detection; representation fusion; emotion specialization; domain-specialized experts; adaptive routing

1. Introduction

News spreads fast online. That speed makes automated credibility checks a critical NLP problem. Arabic adds significant complexity. The language spans Modern Standard Arabic and many regional varieties. Credibility cues shift too—politics looks different from public health or conflict zones. Vocabulary changes. Writing style shifts. Topic distribution varies. Source characteristics differ. These factors create substantial domain shifts between training and evaluation data [1], [2], [3], [4]. A single global classifier struggles under these conditions. It is forced to learn a compromise boundary across heterogeneous distributions. Consequently, domain-specific patterns get lost.

Arabic Transformer encoders like AraBERT [5] and MARBERT [6] offer contextual representations adapted to Arabic. They form strong foundations for detection tasks [7]. Al-Zahrani and Al-Yahya [8] explored ensembles of pretrained models. Their work motivates comparing individual encoders against prediction combinations.

Hybrid approaches mix Transformer representations with other linguistic information [9], [10], [11]. Dahou et al. [9], combined fake-news representations with named entity recognition (NER). Improvements followed on several datasets. Such methods ask a simple question: do different feature spaces offer more than a single encoder?



Emotion is an additional indicator for false information detection. Previous studies looked at emotional signals in news and user reactions [12], [13]. Al-Saif and Al-Dossari [14] examined source and audience emotions in Arabic rumor detection. Explicit features didn't consistently improve strong contextual models. MARBERT can specialize in affective tasks [15]. This gap motivates testing an emotion-specialized representation against veracity-only baselines and generic contexts.

Content heterogeneity remains a hurdle. Domains differ significantly in terminology, semantics, and style [1], [2]. Mixture-of-Experts architectures address this. They maintain specialized subnetworks. A gating mechanism decides their contribution based on input [1],[16]. Recent frameworks adopt this principle. MERMAID uses domain-specialized detectors [3]. MEDAG employs multiple experts with a domain-adaptive gate [4]. Can domain-specialized experts, routed adaptively, beat a single global model? In such a framework, each expert learns domain patterns. Routing dynamically weighs their contributions.

Inspired by these results, we propose EmoFuse. It combines domain-specialized experts, representation-level fusion, and adaptive routing. We don't assume emotion helps. Instead, we explicitly compare emotion-specialized representations against generic MARBERT. The study investigates both adaptive domain specialization and the actual value of emotion-aware information.

The key contribution of this work is summarized below:

- **EmoFuse Framework:** A MARBERT-based multi-expert system for Arabic fake-news detection, using adaptive routing.
- **Representation Fusion:** Combines 128-dim veracity vectors with 768-dim emotion representations into 896-dim vectors per expert.
- **Ablation Study:** Compares emotion-specialized against generic contextual representations to isolate benefits.
- **Routing Evaluation:** Tests domain-supervised soft routing against uniform averaging, hard routing, and a diagnostic oracle.

2. Related Work

Arabic fake-news detection has moved past old-school machine learning. Handcrafted features are dead. Transformers rule now, capturing context and semantics far better than anything before them. AraBERT [5] and MARBERT [6] sit at the heart of this shift. MARBERT, trained on massive social-media data, handles dialectal, informal Arabic much more effectively. That matters when misinformation spreads online. Nassif et al. [7] tested several contextualized language models, comparing AraBERT, ARBERT, MARBERT, AraELECTRA, and QARiB. Results? Contextual representations work well here. Other studies dug into Arabic Transformers for rumor detection [17]. Al-Zahrani and Al-Yahya [8] tried ensemble learning, combining multiple pretrained Arabic models to hit an F1-score of 0.94. These findings prove Arabic-specific Transformers are strong foundations. But there's a catch. Most rely on a single global encoder or use fixed model aggregation. Neither explicitly exploits domain-specialized knowledge. Direct numerical comparison remains tricky because datasets differ. So do label definitions. Preprocessing procedures vary wildly. Evaluation protocols aren't consistent either.

Representation fusion mixes heterogeneous feature spaces before classification. Dahou et al. [9] did exactly that. They combined a 128-dimensional fake-news representation with a 768-dimensional NER one. Their pipeline included feature selection and downstream classification. This motivates investigating auxiliary task specialization. But it doesn't prove every auxiliary task helps. Related work explored linguistic feature concatenation [10]. Others paired AraBERT with graph neural networks [11]. Note the difference: representation fusion combines intermediate features. Ensembles combine model outputs. EmoFuse uses both operations. At distinct stages

Emotion-aware approaches examine affective information to enhance detection capabilities. Guo et al. [12] modeled both publisher and social emotions, while Zhang et al. [18] later investigated dual-emotion representations. FakeFlow [19] tracked how affect changes across news content, and Hamed et al. [13] merged news sentiment with user comment emotions. For Arabic rumor detection, Al-Saif and Al-Dossari [14] tested MARBERT and AraBERT-Twitter using emotional, content, and user features. Notably, explicit emotion features did not consistently boost performance in their study. El-Sayed et al. [15] further investigated Arabic emotion classification with contextual models. Together, these studies suggest that emotion specialization is worth testing, but its usefulness must be evaluated experimentally. A gain from adding a second contextual vector alone does not prove that emotion specialization is the cause

Fake-news characteristics vary significantly by domain, including politics, health, finance, conflict, and entertainment. This variability motivated the development of domain-aware and Mixture-of-Experts (MoE) approaches. MDFEND [1] introduced multiple expert networks where a domain gate adaptively aggregated their representations, establishing an influential architecture for multi-domain detection. FuDFEND [19] relaxed rigid domain assignments using fuzzy modeling, while M3FEND [20] exploited semantic, emotional, and stylistic views alongside domain memory. Collaborative MoE approaches [16] also exist, as do multimodal models like MMDFND [21]. Topic-guided architectures such as TG-MFEND [2] demonstrate value in modeling shared and specific information. TG-MFEND reports that mixing samples from heterogeneous domains creates conflicts, suggesting that explicitly designed multi-domain models exploit shared knowledge better than uniform aggregation. Specialized experts and learned routing generally outperform simple averaging, though expert usefulness varies by domain.

Multi-domain learning poses distinct evaluation conditions. In observed-domain settings like MDFEND, evaluation domains appear during development. MERMAID [3] evaluates transfer to unseen domains using zero-shot experiments, while few-shot adaptation is studied when limited target-domain data becomes available. Its pipeline trains source-specific detectors, adapts and merges models across source domains, and forms predictions via a mixture. MEDAG [22] studies domain-adaptive expert routing under unseen-domain protocols. These works motivate expert combination, yet their transfer settings differ from the observed-domain routing examined here. Consequently, existing experiments do not prove Arabic unseen-domain generalization for EmoFuse.

EmoFuse combines specialized veracity and emotion representations handled by domain-specific Arabic fake-news experts. An input-dependent gate aggregates expert probabilities, while a generic MARBERT representation controls for the addition of a second contextual space. The study focuses on routing across observed domains. Evaluating an entirely unseen domain requires excluding it from every component’s development.

Table 1. Summary of Representative Related Work.

Works	Datasets / Language	Representation	Fusion / Expert Strategy	Finding / Limitation
Nassif et al. [7]	Arabic FND corpora / Arabic	Arabic contextual encoders	Individual encoder classifiers	Strong contextual baselines; no explicit domain routing.
Al-Zahrani and Al-Yahya [8]	Arabic multi-source data / Arabic	Multiple Arabic PLMs	Prediction ensemble	Studies model aggregation; differs from input-dependent domain routing.
Dahou et al. [9]	Seven FND datasets / Arabic	FND128 + NER768	Latent fusion and feature selection	Motivates task-specialized fusion; NER evidence does

				not establish an emotion benefit.
Guo et al. [12]	Weibo / Chinese	Publisher and social emotions	Emotion-aware fusion	Uses affective signals from content and reactions; not Arabic.
FakeFlow [23]	News datasets / English	Semantic and affective information	Affective sequence modeling	Models changes in affect across content; no Arabic domain routing.
Al-Saif and Al-Dossari [14]	Arabic social media / Arabic	Contextual, emotional, and user features	Feature integration	Explicit emotion features do not consistently improve contextual models.
MDFEND [1]	Weibo21 / Chinese	Multiple neural experts	Domain gate and feature aggregation	Motivates domain-aware aggregation; not independently fine-tuned Arabic specialists.
FuDFEND [19]	Weibo21 and Thu / Chinese	Contextual features and fuzzy domain membership	Fuzzy domain gate	Infers non-discrete domain membership; differs from dataset - specific specialists.
Zhao et al. [16]	Multi-domain social-media data	Multiple expert representations	Collaborative MoE	Combines expert knowledge; Arabic emotion specialization is not its focus.
M MDFEND [21]	Multimodal multi-domain data	Text, images, and domain representations	Multimodal expert aggregation	Uses additional modalities; not directly comparable to text-only EmoFuse.
TG-MFEND [2]	Multi-domain news data	Semantic, stylistic, and topic features	Topic-guided multi-view fusion	Models domain differences; specialized Arabic encoders are not its focus.
MERMAID [3]	GossipCop, PolitiFact, MD, CoAID, LIAR / English	Source-specific detectors	Source adaptation, merging, and MoE	Targets unseen domains; few-shot experiments use target data.
MEDAG [4]	Weibo21 and FineFake / Chinese and English	Source-domain experts	Domain-adaptive gate	Studies a stronger transfer setting than observed-domain EmoFuse.

EmoFuse (ours)	AFND and Arabic Covid19 / Arabic	FND128 + EMO768; generic controls	Latent fusion and domain-supervised routing	Routing improves over uniform averaging on Dev; emotion-specific gains are not robust.
-----------------------	---	--	--	---

Table 1 summarizes representations, strategies, settings, limits. Two questions drive the comparison: does adaptive routing beat uniform averaging? Does emotion specialization add value beyond generic context? We don't assume benefits just because the architecture differs

3. Background

3.1 Arabic Fake-News Detection

Arabic fake-news detection can be formulated as a supervised text-classification problem. Given an Arabic text x_i and its corresponding veracity label y_i , the objective is to learn a function that predicts whether the input contains fake or real information. In the binary setting considered in this work,

$$y_i \in \{0,1\} \quad (1)$$

where $y_i = 1$ denotes Fake and $y_i = 0$ denotes Real. A detection model estimates the probability

$$P(y_i = 1 | x_i) \quad (2)$$

then converts it to a binary decision via threshold. You must account for morphological richness, orthographic variation, dialectal diversity, and informal usage. AraBERT [5] and MARBERT [6] offer adapted contextual representations. MARBERT matters most when targeting dialectal or informal social-media text. Class imbalance means overall accuracy is incomplete. Macro-F1 helps here, computing F1-score per class then averaging:

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (3)$$

Here, $C = 2$. Macro-F1 gives equal weight to the Fake and Real class scores. It does not imply that performance is independent of the class distribution.

3.2 Contextual and Task-Specialized Representations

Transformers turn input sequences into contextual token representations.

Let

$$H = f_{\theta}(x) = [h_{CLS}, h_1, \dots, h_n] \quad (4)$$

where f_{θ} denotes a pretrained Transformer encoder and h_{CLS} is the contextual representation associated with the special classification token. For base-sized Arabic BERT-family models, this representation typically belongs to

$$h_{CLS} \in R^{768} \quad (5)$$

Pretrained encoders begin as generic feature extractors. Supervised fine-tuning updates parameters, specializing the latent space. Fine-tune for fake news, and you capture distinctions between fabricated and authentic content. Fine-tune for emotion, and you extract affective patterns. This process motivates the use of task-specialized representations. Two independent copies of the same Transformer, optimized for different objectives, may encode distinct aspects of a single input. However, differing objectives do not guarantee complementarity; redundancy remains a possibility. Therefore, prove the combination's usefulness experimentally rather than assuming it.

3.3 Emotion-Aware Representation Learning

Encode affective information tied directly to text content. Depending on the specific task, represent emotions via sentiment polarity, categorical labels, continuous dimensions, or multi-label annotations. For multi-label problems with categories, map the encoder representation to logits:

$$o^{emo} = W^{emo}h + b^{emo} \quad (6)$$

followed by element-wise sigmoid activation:

$$p^{emo} = \sigma(o^{emo}) \in [0,1]^L \quad (7)$$

Unlike a single-label softmax classifier, sigmoid activation allows several emotions to be associated with the same text simultaneously.

Two forms of information can therefore be extracted from an emotion-specialized model. The first consists of the explicit emotion probabilities

$$p^{emo} \quad (8)$$

which provide a low-dimensional description of predicted emotional categories. The second is the internal contextual representation

$$h_{CLS}^{emo} \quad (9)$$

holds richer affective and semantic details derived from fine-tuning. These components are conceptually distinct. Probabilities reflect predictions, while latent representations retain learned information. A generic encoder might already capture some emotional signal without supervision. Consequently, compare specialized versus generic contexts. Determine whether benefits stem from specialization itself or merely from adding another feature space.

3.4 Representation-Level Fusion

Combine information from multiple sources or learned representations. Methods operate at various levels within the pipeline. Representation-level fusion merges latent vectors before the final classifier. Given two representations,

$$z_1 \in R^{d_1} \quad \text{and} \quad z_2 \in R^{d_2} \quad (10)$$

a simple concatenation strategy produces

$$z^{fusion} = [z_1; z_2] \in R^{d_1+d_2} \quad (11)$$

The fused representation is subsequently provided to a classifier:

$$\hat{y} = g(z^{fusion}) \quad (12)$$

This approach contrasts with decision-level fusion, where models predict independently before results are aggregated. Representation fusion follows:

$$z_1, z_2 \rightarrow [z_1; z_2] \rightarrow \text{Classifier} \quad (13)$$

whereas decision fusion operates as

$$p_1, p_2, \dots, p_K \rightarrow \text{Aggregation} \rightarrow p_{final} \quad (14)$$

The distinction matters. The framework employs both principles but at different stages. Inside an expert, combine latent representations. Across experts, an adaptive routing mechanism merges predictions.

Representation fusion might exploit complementary information. However, higher dimensionality does not guarantee better classification. Redundant vectors add little value and simply increase estimation variance. Thus, single-representation baselines are necessary. Only then can you assess the auxiliary vector's true contribution.

3.5 Domain-Specialized Experts and Adaptive Routing

A Mixture-of-Experts (MoE) framework decomposes a prediction problem into several specialized expert models and a routing or gating mechanism that determines how much each expert contributes to the final prediction.

Let

$$E_1, E_2, \dots, E_K \quad (15)$$

denote K expert models. Given an input x , expert E_k produces a prediction

$$p_k = P(y = 1 \mid x, E_k) \quad (16)$$

A gating network computes a score for each expert and converts these scores into normalized weights through a softmax function:

$$\alpha = \text{softmax}(g(x)) \quad (17)$$

with

$$\alpha_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \alpha_k = 1 \quad (18)$$

In soft routing, the final probability is obtained by weighted aggregation:

$$p_{\text{soft}} = \sum_{k=1}^K \alpha_k p_k \quad (19)$$

Consequently, several experts may contribute simultaneously to one prediction.

By contrast, hard routing selects only the expert receiving the largest routing weight:

$$k^* = \arg \max_k \alpha_k \quad (20)$$

and the final prediction becomes

$$p_{\text{hard}} = p_{k^*} \quad (21)$$

A simple reference strategy is uniform aggregation, where all experts receive identical weights:

$$\alpha_k = \frac{1}{K} \quad (22)$$

Thus,

$$p_{\text{uniform}} = \frac{1}{K} \sum_{k=1}^K p_k \quad (23)$$

Comparing adaptive routing against uniform aggregation reveals something critical. Does the gate actually select experts based on the input? Or does it merely benefit from having multiple models?

Domain-specialized experts capture patterns tied to specific content distributions. The assumption is not that one expert dominates all scenarios. Rather, relevance shifts depending on the input domain. The routing mechanism attempts to identify which specialized knowledge aids most for each prediction.

3.6 Multi-Domain Learning and Cross-Domain Generalization

The terms multi-domain learning, domain adaptation, and domain generalization describe related but distinct experimental settings.

In multi-domain learning, several domains are represented during training:

$$D_{train} = \{D_1, D_2, \dots, D_K\} \quad (24)$$

Evaluation occurs on samples from those observed distributions. The goal is to exploit shared and domain-specific information while minimizing interference. Domain-aware expert routing fits naturally here. Consider MDFEND and related approaches [1], [2].

Domain adaptation works differently. Train on source domains first, then adapt to a target. Depending on the protocol, you may have labeled, unlabeled, or limited target data.

Domain generalization is stricter. Models train exclusively on source domains,

$$D_{source} = \{D_1, \dots, D_K\} \quad (25)$$

while the target domain

$$D_{target} \notin D_{source} \quad (26)$$

is completely excluded from model development. Performance is then measured on this previously unseen domain. Approaches such as MERMAID and MEDAG investigate this stronger form of cross-domain generalization [3], [4].

High domain-classification performance shows that a gate can distinguish observed data distributions. It does not, by itself, demonstrate improved veracity classification or generalization to a new domain. These outcomes must be evaluated separately.

Based on these concepts, the proposed framework combines two levels of adaptation. First, task-specialized contextual representations are fused within each fake-news expert. Second, multiple domain-specialized experts are combined through an input-dependent routing mechanism. The following section describes the resulting EmoFuse architecture and its training procedure in detail.

remains completely out of development. Performance is measured on this unseen territory. MERMAID and MEDAG investigate this stronger form of cross-domain generalization [3], [4].

High domain-classification performance proves a gate can distinguish distributions. That is all. It does not prove improved veracity classification, nor does it guarantee generalization to new domains. Evaluate those aspects separately.

Based on these concepts, the proposed framework combines two levels of adaptation. First, task-specialized contextual representations fuse within each fake-news expert. Second, multiple domain-specialized experts combine through input-dependent routing. The next section details the resulting EmoFuse architecture and its training procedure.

4. Proposed Method

4.1 Framework Overview

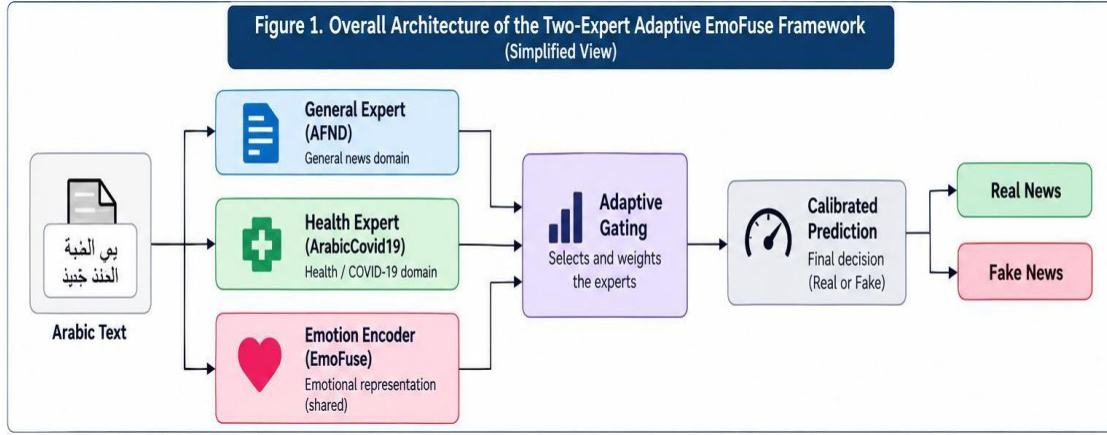


Figure 1. Overall Architecture of the Two-Expert Adaptive EmoFuse Framework.

Figure 1 presents EmoFuse, a framework for Arabic fake-news detection. It combines domain-specialized veracity representations, an independently learned emotion representation, and adaptive expert routing.

At the representation level, each domain expert mixes two elements: a compact veracity representation learned specifically for fake-news classification, plus a contextual representation from an independently emotion-specialized Arabic Transformer. At the decision level, a routing network dynamically weighs each expert’s contribution based on the input text.

The implementation contains two veracity experts. The General expert trains on AFND, while the Health expert trains on ArabicCovid19. A third, independent MARBERT encoder supplies the same emotion representation to both fusion classifiers and the gate. Domain labels match dataset membership. Because dataset identity and thematic domain are coupled here, high routing performance might reflect source, style, or collection differences, not just topic.

Given an Arabic input text $\mathcal{X}$, the overall processing pipeline can be summarized as

$$E_G^{FND}(\mathcal{X}) \rightarrow z_G^{FND} \quad E_H^{FND}(\mathcal{X}) \rightarrow z_H^{FND} \quad E^{EMO}(\mathcal{X}) \rightarrow z^{EMO} \quad (27)$$

followed by expert-specific fusion,

$$z_G^{EF} = [z_G^{FND}; z^{EMO}]$$

$$z_H^{EF} = [z_H^{FND}; z^{EMO}] \quad (28)$$

and adaptive decision fusion,

$$p_{soft} = \alpha_G p_G + \alpha_H p_H \quad (29)$$

4.2 Domain-Specialized Veracity Experts

Veracity experts employ the same MARBERT architecture and pretrained initialization, yet their trainable parameters remain strictly separate. Each encoder, projection layer, and classification head fine-tunes independently on its specific domain. This parameter separation does not imply statistical independence, nor does it guarantee complementary representations.

For each domain k , where

$$k \in \{G, H\} \quad (30)$$

MARBERT transforms the input text $\mathcal{X}$ into contextual token representations:

$$H_k^{FND} = E_k^{FND}(x) = [h_{CLS}^k, h_1^k, \dots, h_n^k] \quad (31)$$

The contextual representation associated with the classification token is

$$h_{CLS}^k \in R^{768} \quad (32)$$

Rather than directly using this full representation for fusion, a trainable projection layer compresses it into a task-oriented latent space:

$$z_k^{FND} = \tanh(W_k^{proj} h_{CLS}^k + b_k^{proj}) \quad (33)$$

where

$$W_k^{proj} \in R^{128 \times 768} \quad (34)$$

resulting in

$$z_k^{FND} \in R^{128} \quad (35)$$

During fake-news specialization, this representation is passed to a binary classification head:

$$s_k = W_k^{cls} z_k^{FND} + b_k^{cls} \quad (36)$$

and

$$p_k^{FND} = \sigma(s_k) \quad (37)$$

The model is optimized using binary cross-entropy:

$$L_{FND}^{(k)} = -[y \log p_k^{FND} + (1 - y) \log(1 - p_k^{FND})] \quad (38)$$

The label convention is

$$y = \{1, Fake, 0, Real\} \quad (39)$$

Crucially, the representation retained for subsequent fusion is the 128-dimensional latent vector extracted before the final neural classification layer, rather than the scalar fake-news probability. This approach allows the fusion stage to leverage richer, task-specialized information.

Once an expert is selected based on development-set performance, it is frozen. The transformer encoder and projection layer cease updating, serving exclusively for feature extraction.

4.3 Emotion-Specialized Representation

A third, independent MARBERT encoder handles multi-label emotion classification, sharing zero trainable parameters with the veracity encoders.

For the same input text x , the emotion encoder produces

$$H^{EMO} = E^{EMO}(x) = [h_{CLS}^{EMO}, h_1^{EMO}, \dots, h_n^{EMO}] \quad (40)$$

with

$$h_{CLS}^{EMO} \in R^{768} \quad (41)$$

During emotion specialization, the contextual representation is mapped to L emotion logits:

$$o^{EMO} = W^{EMO} h_{CLS}^{EMO} + b^{EMO} \quad (42)$$

followed by element-wise sigmoid activation:

$$p^{EMO} = \sigma(o^{EMO}) \quad (43)$$

In our emotion model,

$$L = 28 \quad (44)$$

covers 27 emotion categories plus neutral.

We utilize BCEWithLogitsLoss for this stage. The per-instance expression sums binary cross-entropy over emotion labels; numerically, it operates directly on logits:

$$L_{EMO} = -\sum_{j=1}^L [y_j \log p_j + (1 - y_j) \log(1 - p_j)] \quad (45)$$

The entire emotion encoder fine-tunes during this phase. Once we select a checkpoint using development data, we freeze it. Only then do we extract fake-news representations and train the fusion mechanism.

For the main EmoFuse setup, we retain the final contextual representation

$$z^{EMO} = h_{CLS}^{EMO} \in R^{768} \quad (46)$$

instead of relying solely on the 28 predicted probabilities. This latent vector preserves information from emotion fine-tuning, including context that is not strictly emotional.

4.4 EmoFuse Representation-Level Fusion

Each domain expert receives two independently specialized representations of the same input:

1. a 128-dimensional veracity representation

$$z_k^{FND} \in R^{128} \quad (47)$$

and

2. a 768-dimensional emotion-specialized representation

$$z^{EMO} \in R^{768} \quad (48)$$

They are concatenated to construct the EmoFuse feature vector:

$$z_k^{EF} = [z_k^{FND}; z^{EMO}] \quad (49)$$

Therefore,

$$z_k^{EF} \in R^{896} \quad (50)$$

since

$$128 + 768 = 896 \quad (51)$$

The emotion representation remains shared across experts, while the 128-dim veracity component varies, derived from different domain-specific FND encoders.

Thus, for the two-expert implementation,

$$z_G^{EF} = [z_G^{FND}; z^{EMO}] \quad (52)$$

and

$$z_H^{EF} = [z_H^{FND}; z^{EMO}] \quad (53)$$

Before classification, each fused feature space is standardized using

$$\tilde{z}_k^{EF} = \frac{z_k^{EF} - \mu_k}{\sigma_k} \quad (54)$$

Estimate mean and standard deviation feature-by-feature, utilizing only the relevant domain's training partition. Apply this fitted transform unchanged to development, calibration, and test features.

$$p_k = \sigma(w_k^T \tilde{z}_k^{EF} + b_k) \quad (55)$$

Therefore, the General and Health experts return respectively

$$p_G = P(y = 1 | x, E_G) \quad (56)$$

and

$$p_H = P(y = 1 | x, E_H) \quad (57)$$

This constitutes representation-level late fusion. Transformers specialize independently first; subsequently, we combine their frozen latents for downstream fake-news classification.

4.5 Adaptive Expert Routing

Uniform averaging assigns equal weight to every expert. In contrast, EmoFuse learns an input-dependent domain distribution, using those probabilities as weights. Domain membership serves as a proxy for expert suitability, rather than providing a direct error estimate.

For the primary architecture, the gate utilizes the shared emotion-specialized contextual representation

$$z^{EMO} \in R^{768} \quad (58)$$

as its input.

The gate consists of a fully connected hidden layer:

$$h^{gate} = \text{ReLU}(W_1^{gate} z^{EMO} + b_1^{gate}) \quad (59)$$

with

$$h^{gate} \in R^{128} \quad (60)$$

The hidden representation is passed through dropout with probability 0.1 and a linear layer with two outputs. Dropout is disabled at inference:

$$a = W_2^{gate} \text{Dropout}_{0.1}(h^{gate}) + b_2^{gate} \quad (61)$$

A softmax converts these logits into normalized expert weights:

$$[\alpha_G, \alpha_H] = \text{softmax}(a) \quad (62)$$

such that

$$\alpha_G + \alpha_H = 1 \quad (63)$$

and

$$\alpha_G, \alpha_H \geq 0 \quad (64)$$

Train the gate on domain labels, not Fake/Real labels. Its objective is to predict the instance’s origin, using the resulting probabilities as weights. Note that this does not guarantee the chosen expert is most accurate for every single case.

For an instance originating from domain d , the routing loss is

$$L_{gate} = - \sum_{k=1}^K d_k \log \alpha_k \quad (65)$$

where d_k represents the domain target.

Veracity experts learn Fake/Real discrimination, while the gate learns domain identification. Report routing improvements and final classification gains separately.

4.6 Final Prediction

Soft routing serves as our primary aggregation strategy.

For two experts, the raw final fake-news probability is

$$p_{soft} = \alpha_G p_G + \alpha_H p_H \quad (66)$$

For comparison, two alternative aggregation strategies are considered.

Both contribute, weighted by the estimated domain distribution.

Uniform aggregation

$$p_{uniform} = \frac{p_G + p_H}{2} \quad (67)$$

Does learned routing outperform simple multi-expert averaging?

Hard routing

The selected expert is

$$k^* = \arg \max_{k \in \{G, H\}} \alpha_k \quad (68)$$

and

$$p_{hard} = p_{k^*} \quad (69)$$

This offers a direct contrast to soft weighting.

The domain oracle selects the expert matching the known label. Use it strictly for diagnostics. It is not an upper bound—other experts or combinations might perform better on specific instances. The oracle assumes true labels; deployed models must infer them

4.7 Probability Calibration

Combining predictions from separately trained experts means raw probabilities are not perfectly calibrated.

Calibrate on a separate partition after locking architecture and hyperparameters via development data.

Let

$$p_{soft} \quad (70)$$

denote the raw routed probability. For probabilities strictly between zero and one, its log-odds are

$$l = \log \frac{p_{soft}}{1 - p_{soft}} \quad (71)$$

Platt scaling then learns two calibration parameters a and b :

$$p_{cal} = \sigma(al + b) \quad (72)$$

The final classification threshold τ is selected exclusively on the calibration partition:

$$\hat{y} = \{1, p_{cal} \geq \tau, 0, p_{cal} < \tau. \quad (73)$$

Fit both calibrator and threshold using only calibration data. We evaluated the official test set earlier ; keep that result historical. Later development comparisons do not count as new untouched test confirmation

4.8 Training Procedure

Training proceeds sequentially. Fine-tune encoders during specialization, then freeze them while training fusion classifiers and the domain gate.

Stage 1 — Emotion specialization

An independent MARBERT model is fine-tuned on a multi-label Arabic emotion-classification task:

$$MARBERT \rightarrow^{emotion\ fine-tuning} E^{EMO} \quad (74)$$

After checkpoint selection, E^{EMO} is frozen.

Stage 2 — Domain-specific fake-news specialization

A separate MARBERT copy is initialized for each fake-news domain:

$$MARBERT_G \rightarrow^{DG} E_G^{FND}$$

$$MARBERT_H \rightarrow^{DH} E_H^{FND} \quad (75)$$

Each encoder is fully fine-tuned with its 768-to-128 projection and binary classification head.

After model selection, the domain experts are frozen.

Stage 3 — Representation extraction

For every fake-news instance, the framework extracts

$$z_G^{FND}, z_H^{FND}, z^{EMO} \quad (76)$$

The emotion representation is computed once and can be reused by all domain experts.

Stage 4 — Expert-level fusion

For each expert,

$$z_k^{EF} = [z_k^{FND}; z^{EMO}] \in R^{896} \quad (77)$$

A StandardScaler and regularized logistic regression are trained using only the corresponding domain training partition.

Stage 5 — Gate training

The routing network is trained on the combined multi-domain training data using domain labels:

$$z^{EMO} \rightarrow Gate \rightarrow [\alpha_G, \alpha_H] \quad (78)$$

The gate checkpoint is selected on the development partition using domain-routing performance. This selection criterion is distinct from the final Fake/Real Macro-F1.

Stage 6 — Adaptive inference

At inference time, the same input is passed through the two frozen domain experts and the shared emotion encoder:

$$x \rightarrow p_G, p_H, \alpha_G, \alpha_H \quad (79)$$

The raw final probability is

$$p_{soft} = \alpha_G p_G + \alpha_H p_H \quad (80)$$

Stage 7 — Calibration and final decision

Finally,

$$p_{soft} \xrightarrow{\text{logit } l} \xrightarrow{\text{Platt scaling}} p_{cal} \xrightarrow{\tau} \hat{y} \quad (81)$$

Figure 2 details the complete inference path, including feature standardization, routing, and calibration.

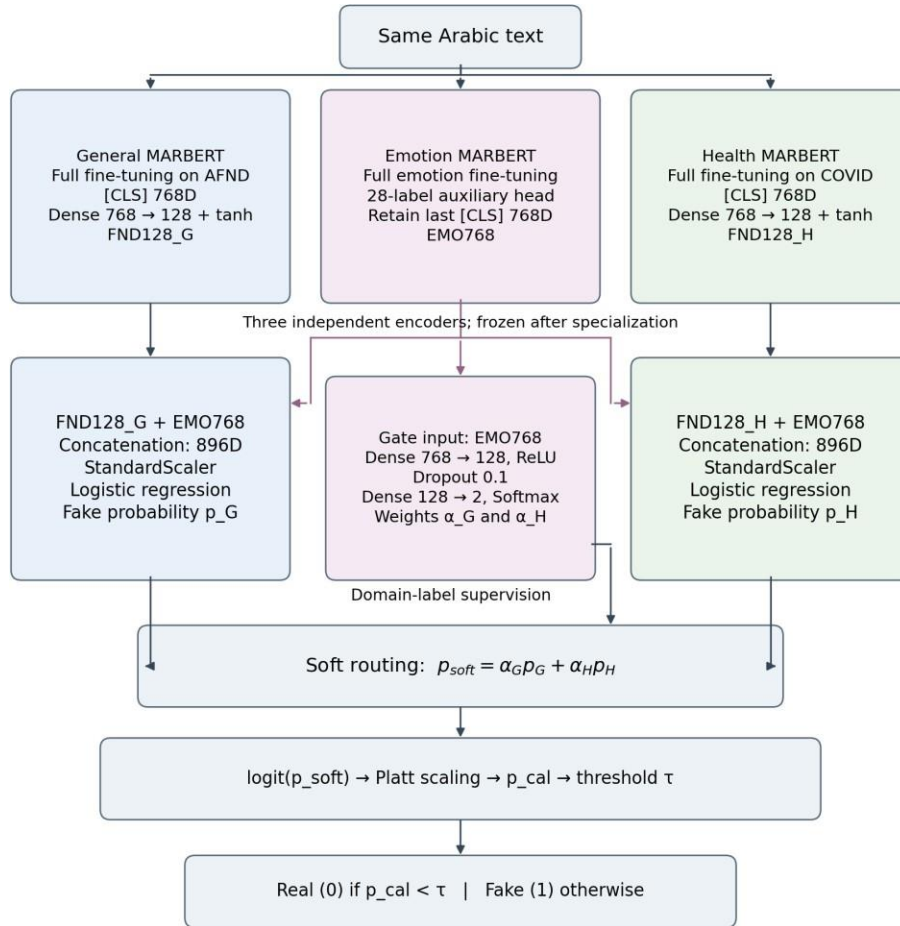


Figure 2. Detailed technical architecture of the proposed framework EmoFuse.

4.9 Control Configurations

Did gains come from emotion specialization? Or just adding another context layer? We tested two controls to find out. First: FND-only. Just the 128-dimensional veracity rep. Second: GenericFuse. Swapped the emotion-specialized MARBERT for a 768-dim version from original generic MARBERT—no fine-tuning:

$$z_k^{GenericFuse} = [z_k^{FND}; z_k^{Generic}] \quad (82)$$

These controls enable three representation conditions to be compared:

$$FND128$$

$$FND128 + EMO768 \quad (83)$$

and

$$FND128 + Generic768 \quad (84)$$

Veracity-only vs. emotion-fused vs. generic-fused. Plus routing checks: uniform averaging, hard routing, diagnostic oracle, and a gate taking Generic768 instead of EMO768. Baseline? A single global MARBERT on combined partitions. Result in dev experiments: Emotion specialization didn't give a robust boost. The main EmoFuse setup? Didn't clearly beat the global baseline either.

5. Experimental

We evaluate EmoFuse on two datasets under observed-domain settings. Three questions drive this: Does adaptive routing beat uniform averaging of the same experts? Does emotion specialization help beyond veracity or generic contexts? And does the whole system top a single global MARBERT?

5.1 Dataset

5.1.1 Dataset Description

Two Arabic public datasets. First: Arabic Fake News Dataset (AFND) [24]. Sourced from various public news sites. 606,912 articles across 134 platforms. Breakdown: 369,130 credible, 167,237 non-credible, 225,512 undecided. We ignore the undecided ones for this research. AFND acts as the General dataset. Second: Arabic Covid19 [25]. Twitter posts on COVID-19. Jan 1, 2020 to May 31, 2020. 34,552 tweets total. 15,168 real, 19,384 fake.

5.1.2 Data Processing

Data processing is a crucial phase in creating effective model [26] include orthographic correction, normalization, deduplication. Normalization steps: map Alef variants to plain Alef; strip diacritics; collapse repeated characters; mask URLs and numbers. Deduplication keeps one copy per identical text-label pair. Conflicting labels on duplicates? Addressed to keep annotation consistent.

5.1.3 Leakage Controlled Partitions

AFND and Arabic Covid19 (COVID) are used, respectively, for the general and health domains.

After cleaning AFND, we take a fixed stratified sample of 30,000 from AFND. Cleaned COVID has 34,864 tweets. Labels are binary: 0 = Real, 1 = Fake. No reliable gold event IDs existed. So we built leakage groups from near-duplicates and similar texts. Cross-domain split: 70% train, 10% dev, 10% calib, 10% test. The split is realised by the leakage group, not by individual instances. All group members land in the same bucket.

Partitions ended up group-disjoint and instance-disjoint. It reduces leakage from similar variants. Doesn't guarantee all texts about the same event stay together, though. Table 2 and figure3 show these fixed splits used below.

Table 2. Experimental partition sizes after leakage-group construction.

Dataset	Train	Dev	Calibration	Test	Total
AFND	21,002	3,000	2,999	2,999	30,000
COVID	24,406	3,486	3,486	3,486	34,864

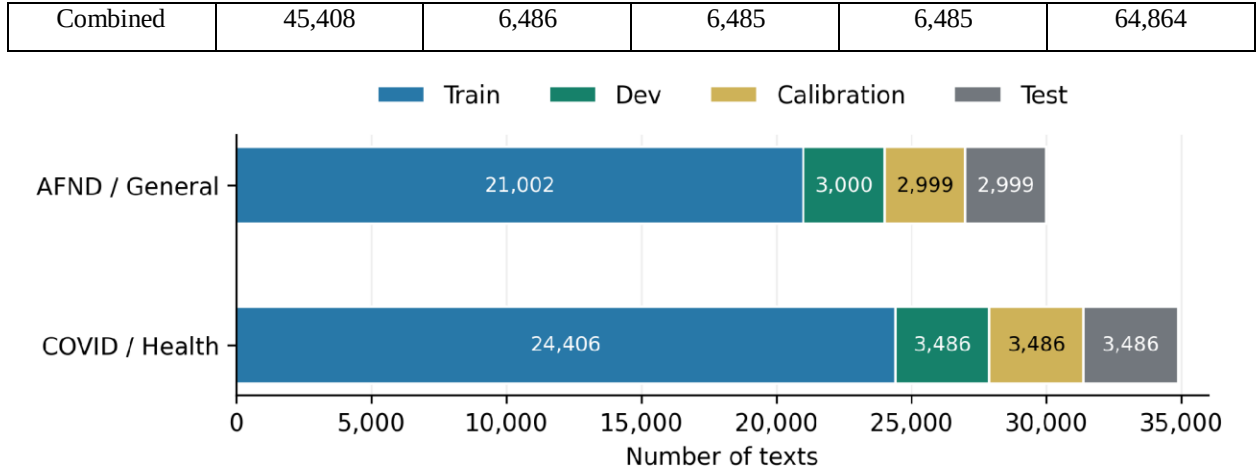


Figure 3. Distribution of texts across the four group-disjoint partitions (70/10/10/10 split).

5.2. Experimental Setup

Independent MARBERT experts [6] use 768-to-128 projections. Tanh activation, dropout, binary heads. Fine-tuned fully for three epochs. Adamax optimizer. Learning rate: 5×10^{-5} . Micro-batches of one. 32 gradient accumulation steps. Effective batch size hits 32. FP16 mixed precision enabled. Gradient checkpointing too. Max sequence length: 256 tokens for AFND. 128 for COVID.

The auxiliary emotion encoder? Fine-tuned independently on the Arabic subset of go_emotions. We used binary cross-entropy with logits across 28 labels, picking the best model on the emotion development partition (Section 4.3). Once selected, we froze those encoders and cached their features. Both EMO768 and Generic768 took 128-token inputs, using identical texts and extraction order. Note that Generic768 came straight from MARBERT, no emotion fine-tuning.

For each expert, we fitted a StandardScaler exclusively on its own-domain training features. Then came L2-regularized logistic regression. We selected C values from {0.01, 0.1, 1, 10} on Dev, using either unweighted or balanced class weights at a fixed 0.50 threshold. The gate relied on domain-balanced training samples and domain-label supervision. Checkpoint selection? Based purely on Dev routing performance. Its input was EMO768 in the primary system, but Generic768 in the routing control.

Test comparisons covered individual experts, uniform averaging with equal weights, soft routing with learned weights, hard routing (picking the highest-weight expert), and a dataset-label oracle. That oracle serves as a diagnostic reference, not a guaranteed upper bound. For Dev ablations, we compared FND128, EmoFuse896, and GenericFuse896 under identical partitions and classifier-selection procedures. We also varied the gate input while keeping EmoFuse experts fixed, evaluating veracity-only experts with generic routing.

One global MARBERT baseline was trained on the combined 45,408 training instances. It used 256-token inputs, three epochs, and a learning rate of 2×10^{-5} . Checkpoints were selected by combined Dev macro-F1. We didn't match total parameter count or inference cost across architectures.

Platt scaling and threshold selection used Calibration data exclusively. This yielded a 0.478 threshold for the historical system; uncalibrated comparisons stuck to 0.50. Macro-F1 is our primary veracity metric, backed up by class-wise F1, accuracy, ROC-AUC, Brier score, and log-loss where available. Combined macro-F1 comes from concatenated predictions. Routing metrics are reported separately. Crucially, no calibration parameters or thresholds were fitted on Test.

Dev experiments used seeds 42, 100, and 200 on fixed partitions. We retrained veracity experts and gates but reused frozen auxiliary representations. Results appear as mean $\pm$ standard deviation, with paired differences included where possible.

6. Results

Here's how the proposed EmoFuse framework stacked up. We tested it against single-model baselines, alternative expert-combination strategies, and representation-level ablations. The goal? To see what domain-specialized experts, adaptive routing, and emotion-oriented representation fusion actually contribute.

6.1 Performance Evaluation

Table 4 summarises performance on the 6,485 combined Test texts. The final calibrated soft system hit Macro-F1 = 0.8735, accuracy = 0.8736, and ROC-AUC = 0.9433. Class-wise F1 scores landed at 0.8750 for Fake and 0.8720 for Real. Raw soft routing reached Macro-F1 = 0.8734. Uniform averaging? Only 0.8182. That's an absolute gain of 0.0552 or 5.52 percentage points using the same two expert classifiers.

Table 4. Test performance on the combined 6,485 texts.

System	Macro-F1	Accuracy	ROC-AUC
General only	0.6823	0.6877	0.7630
Health only	0.7047	0.7055	0.7738
Uniform averaging	0.8182	0.8184	0.8684
Soft routing	0.8734	0.8734	0.9433
Hard routing	0.8728	0.8728	0.9427
Domain oracle	0.8765	0.8765	0.9462
Calibrated soft	0.8735	0.8736	0.9433

The domain oracle achieves a Macro-F1 score of 0.8765, resulting in a discrepancy of 0.0031 compared to raw soft routing. Hard routing attains a value of 0.8728, merely 0.0006 less than raw soft routing. The discerned benefit pertains to adaptive expert selection rather than uniform averaging, rather than a significant distinction between soft and hard decisions. Figure 4 indicates that neither expert excels individually throughout the aggregated distribution: General-only and Health-only attain Macro-F1 scores of 0.6823 and 0.7047, respectively.

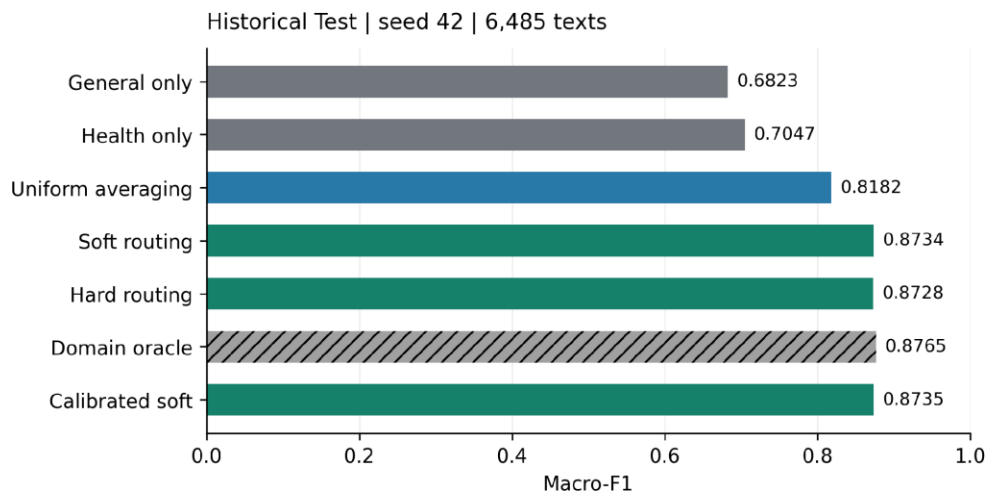


Figure 4. Veracity Macro-F1 on the historical Test set..

6.2 Expert Specialization and Routing Behavior

As shown in Figure 5a the cross domain examination shows a strong specialisation of the experts. The General expert obtained a Macro-F1 of 0.8926 on AFND and 0.4409 on COVID. However, the health specialist obtained a score of 0.8591 for COVID, but this fell to 0.4931 for AFND. These differences call for adaptive weights. Allowing a specialist to work outside of its training dataset reduces overall efficacy. However, the cumulative scores do not indicate if particular predictions are negatively affected.

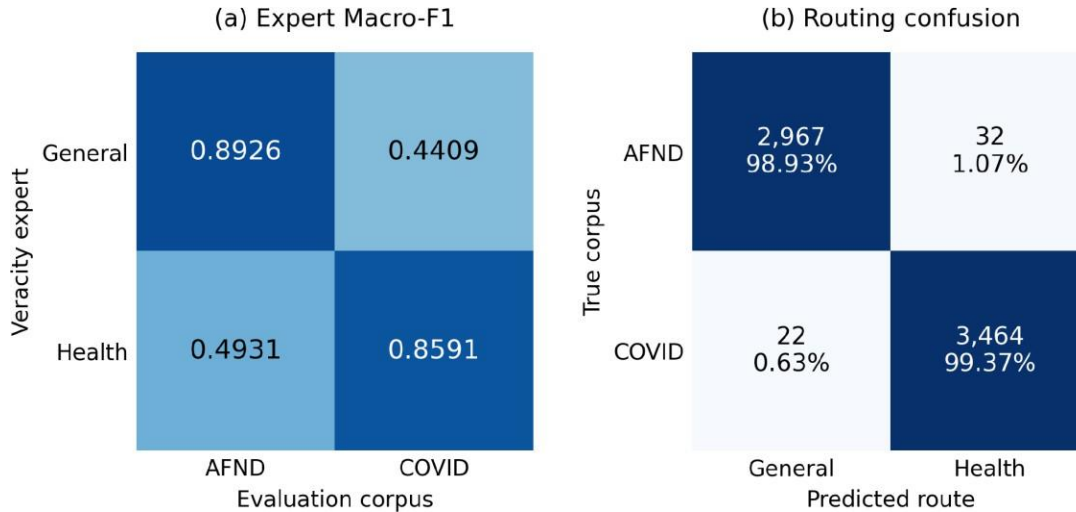


Figure 5. Test diagnostics Panel (a) reports expert performance on each dataset. Panel (b) presents the domain-routing confusion matrix.

The EMO768 gate achieved an impressive routing Accuracy of 0.9917 and a routing Macro-F1 score of 0.9916. It successfully assigned 2,967 out of 2,999 AFND texts to the General category and 3,464 out of 3,486 COVID texts to Health. The total number of routing errors was merely 54 (refer to Figure 5b). The mean weights for the corresponding experts were notably high: 0.9848 for AFND and 0.9893 for COVID. This consistency is reflected in the similar aggregate performance observed in both soft and hard routing. However, it is important to note that a routing error does not necessarily imply a veracity error. The "wrong" expert could still accurately predict the correct Fake/Real label.

The final calibrated system achieves an AFND Test Macro-F1 of 0.8893 and ROC-AUC of 0.9578, compared with 0.8563 and 0.9305, respectively, on COVID. The unweighted mean of the two dataset -level Macro-F1 scores is 0.8728, whereas the Macro-F1 computed from concatenated Test predictions is 0.8735. Although numerically close, these statistics use different aggregation procedures and should not be treated as interchangeable. COVID remains more difficult for veracity classification despite its higher routing accuracy, indicating that accurate dataset assignment alone does not ensure correct Fake/Real predictions.

6.3 Effect of Probability Calibration

Platt calibration tightens probability quality on the historical seed-42 Test set. Look at Figure 6: Brier score drops from 0.1002 to 0.0920, while log-loss falls from 0.3855 to 0.3045. That's a relative reduction of roughly 8.2% and 21.0%, respectively. In comparison, Macro-F1 increases very marginally, from 0.8734 to 0.8735, following calibration and threshold correction.

ROC-AUC remains unchanged at 0.9433, consistent with the strictly increasing Platt transformation preserving the ordering of prediction scores. Because probability calibration and threshold adjustment are applied together, their individual contributions to the small Macro-F1 improvement cannot be isolated from this comparison. The lower Brier

score and log-loss support improved probabilistic performance, but do not establish calibration across all probability ranges.

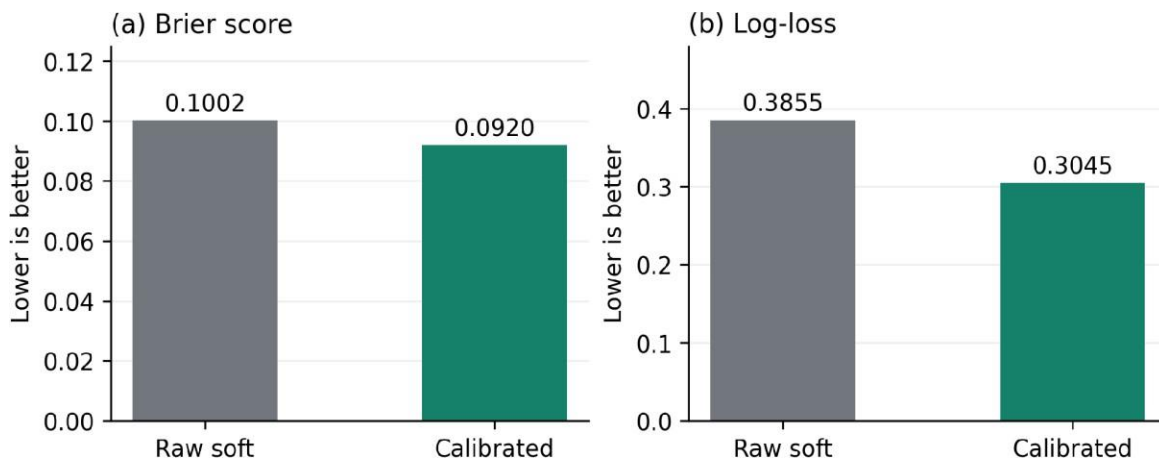


Figure 6. Brier score (a) and log-loss (b) before and after Platt calibration on the historical seed-42 Test set.

6.4 Stability of Adaptive Routing

Adaptive soft routing surpasses uniform averaging across all three training seeds on the fixed Dev partition (Figure 7). Take seeds 42, 100, and 200. Uniform averaging hits Macro-F1 scores of 0.8126, 0.8242, and 0.8238. Compare that to 0.8677, 0.8683, and 0.8679 for the primary soft-routing system. Gains sit around 5.51, 4.41, and 4.41 percentage points. Soft routing averages a Macro-F1 of 0.8680 ± 0.0003 ; uniform averaging manages only 0.8202 ± 0.0066 . The mean paired improvement is 0.04777 ± 0.00635 about 4.78 percentage points. Clearly, adaptive routing offers a consistent advantage with less variability across runs.

The representation-matched EmoFuse domain oracle reaches 0.8713 ± 0.0004 Macro-F1 on Dev. with a recorded paired oracle-minus-soft gap of 0.0034 ± 0.0007 . These numbers show a small remaining loss compared to explicit dataset assignment for these experts. They don't, however, establish an oracle bound for other expert representations or new domains. Three-seed repetitions strengthen the consistency of this uniform-versus-adaptive comparison on this partition. where, do not replace independently held-out confirmation.

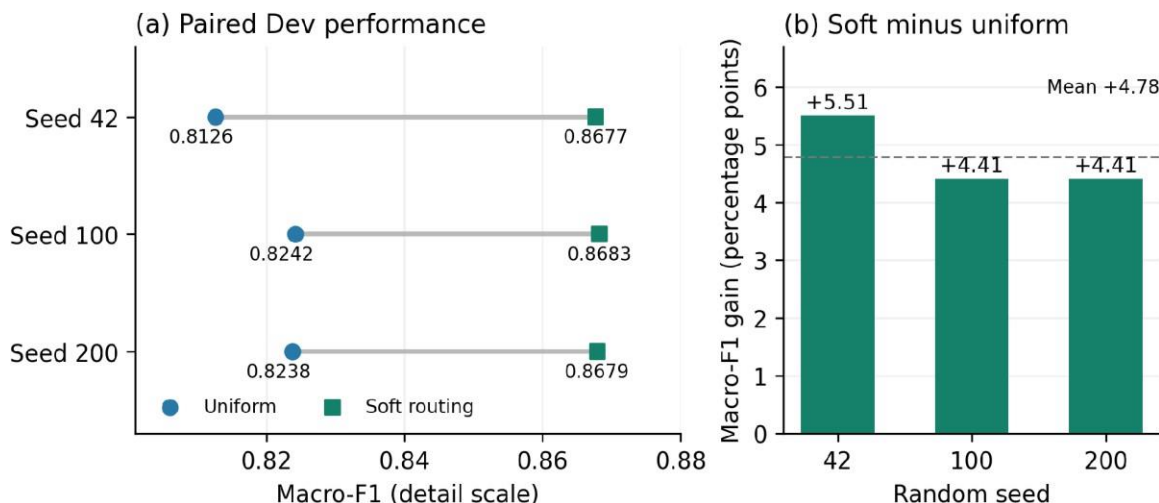


Figure 7. Dev performance across three training seeds. Panel (a) compares uniform averaging and soft routing using the same experts for each seed. Panel (b) shows the corresponding paired gains in percentage points

6.5 Emotion representation and Gate Ablations

Figure 8a looks at expert representations on their corresponding Dev datasets. For General on AFND, mean Macro-F1 is 0.8720 ± 0.0044 for FND128, 0.8692 ± 0.0042 for EmoFuse896, and 0.8686 ± 0.0042 for GenericFuse896. On Health for COVID, scores are 0.8701 ± 0.0039 , 0.8692 ± 0.0030 , and 0.8706 ± 0.0052 . Adding emotion-specialized representation doesn’t improve mean performance over FND128 in either dataset. GenericFuse896 does edge out slightly on COVID, but that advantage isn’t consistent across seeds.

Figure 8b illustrates the performance of combined-domain veracity. Using the same seed-specific EmoFuse expert predictions, substituting the EMO768 gate input with Generic768 results in an increase in mean Macro-F1 from 0.8680 ± 0.0003 to 0.8701 ± 0.0003 . The FND128 experts, employing a Generic768 gate, achieve the highest mean performance among the configurations presented, at 0.8717 ± 0.0014 , in contrast to 0.8671 ± 0.0045 for the global MARBERT model. These comparisons do not indicate a consistent advantage from explicit emotion fine-tuning under the assessed protocol. However, they do not suggest that generic contextual representations lack emotional information.

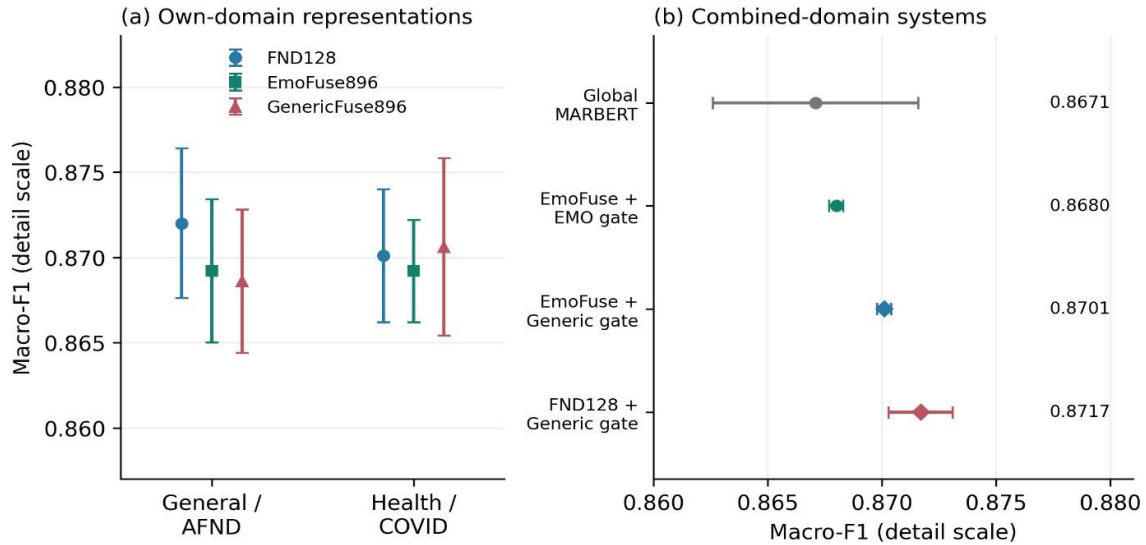


Figure 8. Emotion representation and system ablations . Panel (a) compares own-domain expert representations; panel (b) compares combined-domain systems.

6.6 Global MARBERT Baseline Evaluation

Table 5. Combined Dev comparisons across seeds 42, 100, and 200. Values are mean $\pm$ standard deviation; deltas are paired within seeds and summarized before rounding.

System	Macro-F1	Paired Delta vs global	Status
Global MARBERT	0.8571 ± 0.0031	Reference	Baseline
EmoFuse + EMO Gate	0.8680 ± 0.0003	$+0.0009 \pm 0.0048$	Primary
EmoFuse + Generic Gate	0.8701 ± 0.0003	$+0.0031 \pm 0.0043$	Post-hoc Dev
FND128 + Generic Gate	0.8717 ± 0.0014	$+0.0047 \pm 0.0057$	Post-hoc Dev

The global MARBERT baseline, fine-tuned jointly on AFND and COVID, gets a mean Dev Macro-F1 of 0.8571 ± 0.0031 . Our primary EmoFuse + EMO Gate system hits 0.8680 ± 0.0003 . That’s a mean improvement of

0.0109, or 1.09 percentage points. Under this evaluated protocol, the multi-expert system clearly outperforms the jointly trained baseline on Dev.

The generic-gate configurations attain superior average scores: 0.8701 ± 0.0003 for EmoFuse combined with Generic Gate and 0.8717 ± 0.0014 for FND128 paired with Generic Gate. According to the presented averages, their enhancements over Global MARBERT are 1.30 and 1.46 percentage points, respectively. These configurations persist as post-hoc development analyzes necessitating independent test validation. The noted benefit cannot be exclusively linked to explicit emotional calibration, as their findings further suggest.

The historically calibrated Test macro-F1 score of 0.8735 is solely applicable to the seed-42 primary system. It cannot be allocated to the generic-gate variants nor directly juxtaposed with the global baseline's Dev score.

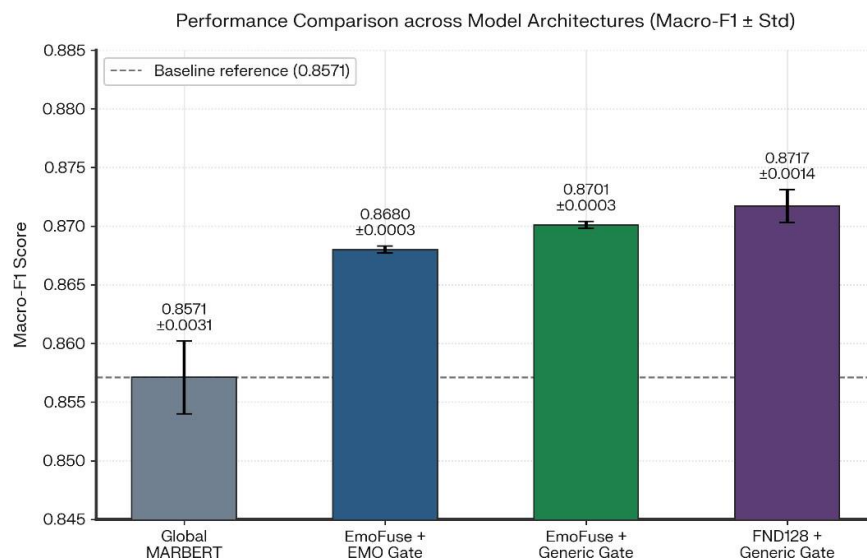


Figure 9. macro-F1 comparison across evaluated model architectures.

8. Discussion

Input-dependent routing helps most when combining strongly dataset-specialized experts. Since uniform averaging and learned routing use the same expert predictions, comparing them reveals how much the aggregation mechanism matters. The 5.52 points historical Test difference aligns with a positive Dev difference for every seed, averaging 4.78 points. Add the cross-dataset diagnostic to that, and a pattern emerges. Routing largely avoids degradation by preventing substantial weight from going to an expert outside its training distribution.

This interpretation is more specific than a general claim that a multi-expert architecture is superior to a single contextual model. Almost identical to the primary system, the robust global baseline achieves a mean development performance. The evidence suggests that an adaptive combination of specialists is more effective than an unsuitable uniform combination of specialists. However, it does not suggest a significant overall advantage of primary EmoFuse over joint MARBERT training. This distinction becomes particularly noteworthy when considering the supplementary complexity introduced by three contextual encoders, distinct classifiers, and a gating mechanism.

The idea of emotion specialization was a testable design proposition, although it did not assure enhancement. Under the stringent group-disjoint methodology, the multi-seed controls fail to demonstrate an additional mean advantage from including EMO768 in FND128. Generic representations provide more precise domain routing compared to emotion-specific representations. Thus, the efficacy of the primary method cannot be construed as proof that its superiority over uniform averaging is influenced by emotional signals.

Why Several explanations fit. Auxiliary and veracity representations are probably redundant. Or emotion fine-tuning strips domain-discriminative information. Maybe the auxiliary dataset differs in genre and annotation from fake-news corpora. These are hypotheses, not tested findings. An EMO768 vector can hold lexical data unrelated to explicit emotion categories. Controls isolate emotion-oriented fine-tuning effects better than architecture comparisons alone. They don't identify a causal mechanism. They prove emotion is relevant in at least one Arabic misinformation setting.

The gate assigns the majority of the weight to the expert who matches the dataset, which explains the numerical similarity between soft and hard routing. A soft mixture primarily functions as a near-discrete choice in this context. It would be premature to assert that soft routing has demonstrated advantages for ambiguous or mixed-domain texts, as these scenarios were not independently assessed. In the same vein, the proximity to the domain oracle indicates that the performance of the same experts is little affected in comparison to dataset-label assignment. However, the optimal expert selection may not be achieved in every instance.

.Platt scaling improves recorded probability scores. Ranking and binary classification stay nearly unchanged. This separates probability quality from classification benefit. Better aggregate Brier score and log-loss don't guarantee reliable confidence for every class, subgroup, or future domain. You'd need reliability plots, subgroup calibration analysis, and prospective threshold assessment before using these probabilities for operational decisions.

The principal limitations of this work are the evaluation of two observed datasets defined-domains only, so high routing accuracy may reflect dataset-specific quirks, not general domain understanding. Official test results rely on a single seed-42 evaluation. We also didn't systematically compare computational costs against baselines.

9. Conclusion

We introduced EmoFuse: a two-expert MARBERT framework for Arabic fake-news detection. It combines domain-specific representations with adaptive expert routing. Using leakage-group-disjoint partitions, the learned gate improved Macro-F1 over uniform averaging. It approached domain-oracle performance.

Calibration boosted probability reliability. Classification performance barely moved. Experiments found no consistent evidence that emotion-specific representations improve detection beyond generic or veracity-oriented ones. Findings apply only to routing between AFND and ArabicCovid19 datasets. Generalization to unseen domains isn't established. Future work needs leave-one-domain-out evaluation. Strongest evidence supports combining domain-specialized experts with adaptive routing. The specific contribution of emotion specialization still unresolved.

REFERENCES

1. Q. Nan, J. Cao, Y. Zhu, Y. Wang, et J. Li, « MDFEND: Multi-domain Fake News Detection », in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Virtual Event Queensland Australia: ACM, oct. 2021, p. 3343-3347. doi: 10.1145/3459637.3482139.
2. L. Wang et Y. Hu, « Topic-guided multi-domain fake news detection », *Multimed. Syst.*, vol. 31, no 2, p. 75, avr. 2025, doi: 10.1007/s00530-024-01636-x.
3. A. Liguori, F. S. Pisani, C. Comito, M. Guarascio, et G. Manco, « Breaking domain barriers: mixture of experts for cross-domain fake news detection », *Mach. Learn.*, vol. 114, no 8, p. 188, août 2025, doi: 10.1007/s10994-025-06827-9.
4. Z. Guo, J. Ouyang, X. Li, et C. Li, « Domain adaptive mixture-of-experts for cross-domain fake news detection », *Knowl.-Based Syst.*, vol. 340, p. 115535, mai 2026, doi: 10.1016/j.knosys.2026.115535.
5. W. Antoun, F. Baly, et H. Hajj, « AraBERT: Transformer-based Model for Arabic Language Understanding », in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, et H. Mubarak, Éd., Marseille, France: European Language Resources Association, mai 2020, p. 9-15. Consulté le: 12 septembre 2026. [En ligne]. Disponible sur: <https://aclanthology.org/2020.osact-1.2/>
6. M. Abdul-Mageed, A. Elmadany, et E. M. B. Nagoudi, « ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic », in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, et R. Navigli, Éd., Online: Association for Computational Linguistics, août 2021, p. 7088-7105. doi: 10.18653/v1/2021.acl-long.551.

7. A. B. Nassif, A. Elnagar, O. Elgendy, et Y. Afadar, « Arabic fake news detection based on deep contextualized embedding models », *Neural Comput. Appl.*, vol. 34, no 18, p. 16019-16032, sept. 2022, doi: 10.1007/s00521-022-07206-4.
8. L. Al-Zahrani et M. Al-Yahya, « Pre-Trained Language Model Ensemble for Arabic Fake News Detection », *Mathematics*, vol. 12, no 18, p. 2941, sept. 2024, doi: 10.3390/math12182941.
9. A. Dahou *et al.*, « Linguistic feature fusion for Arabic fake news detection and named entity recognition using reinforcement learning and swarm optimization », *Neurocomputing*, vol. 598, p. 128078, sept. 2024, doi: 10.1016/j.neucom.2024.128078.
10. M. Lichouri, K. Lounnas, A. Zitouni, H. Latrache, et R. Djeradi, « USTHB at ArAIEval'23 Shared Task: Disinformation Detection System based on Linguistic Feature Concatenation », in *Proceedings of ArabicNLP 2023, Singapore (Hybrid): Association for Computational Linguistics*, 2023, p. 508-512. doi: 10.18653/v1/2023.arabnlp-1.47.
11. O. Karajeh, M. N. Al-Kabi, et E. A. Fox, « Fusing AraBERT and Graph Neural Networks for Enhanced Arabic Text Classification », in *2023 24th International Arab Conference on Information Technology (ACIT)*, déc. 2023, p. 1-8. doi: 10.1109/ACIT58888.2023.10453909.
12. C. Guo, J. Cao, X. Zhang, K. Shu, et M. Yu, « Exploiting Emotions for Fake News Detection on Social Media », 5 mars 2019, *arXiv: arXiv:1903.01728*. doi: 10.48550/arXiv.1903.01728.
13. S. Kh. Hamed, M. J. Ab Aziz, et M. R. Yaakub, « Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments », *Sensors*, vol. 23, no 4, p. 1748, févr. 2023, doi: 10.3390/s23041748.
14. H. F. Al-Saif et H. Z. Al-Dossari, « Exploring the Role of Emotions in Arabic Rumor Detection in Social Media », *Appl. Sci.*, vol. 13, no 15, p. 8815, juill. 2023, doi: 10.3390/app13158815.
15. A. El-Sayed, M. Abougabal, et S. Lazem, « Benchmarking a large Twitter dataset for Arabic emotion analysis », *SN Appl. Sci.*, vol. 5, no 8, p. 224, août 2023, doi: 10.1007/s42452-023-05437-1.
16. J. Zhao, Z. Zhao, L. Shi, Z. Kuang, et Y. Liu, « Collaborative Mixture-of-Experts Model for Multi-Domain Fake News Detection », *Electronics*, vol. 12, no 16, p. 3440, août 2023, doi: 10.3390/electronics12163440.
17. S. Ben Ali, Z. Kechaou, et A. Wali, « Arabic fake news detection in social media Based on AraBERT », in *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Toronto, ON, Canada: IEEE, déc. 2022, p. 214-220. doi: 10.1109/ICCICC57084.2022.10101635.
18. X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, et K. Shu, « Mining Dual Emotion for Fake News Detection », in *Proceedings of the Web Conference 2021*, Ljubljana Slovenia: ACM, avr. 2021, p. 3465-3476. doi: 10.1145/3442381.3450004.
19. C. Liang, Y. Zhang, X. Li, J. Zhang, et Y. Yu, « FuDFEND: Fuzzy-Domain for Multi-domain Fake News Detection », in *Natural Language Processing and Chinese Computing*, vol. 13552, W. Lu, S. Huang, Y. Hong, et X. Zhou, Éd., in *Lecture Notes in Computer Science*, vol. 13552, Cham: Springer Nature Switzerland, 2022, p. 45-57. doi: 10.1007/978-3-031-17189-5_4.
20. Y. Zhu *et al.*, « Memory-Guided Multi-View Multi-Domain Fake News Detection », *IEEE Trans. Knowl. Data Eng.*, p. 1-14, 2022, doi: 10.1109/TKDE.2022.3185151.
21. Y. Tong, W. Lu, Z. Zhao, S. Lai, et T. Shi, « MDMFND: Multi-modal Multi-Domain Fake News Detection », in *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne VIC Australia: ACM, oct. 2024, p. 1178-1186. doi: 10.1145/3664647.3681317.
22. Z. Guo, J. Ouyang, X. Li, et C. Li, « Domain adaptive mixture-of-experts for cross-domain fake news detection », *Knowl.-Based Syst.*, vol. 340, p. 115535, mai 2026, doi: 10.1016/j.knosys.2026.115535.
23. B. Ghanem, S. P. Ponzetto, P. Rosso, et F. Rangel, « FakeFlow: Fake News Detection by Modeling the Flow of Affective Information », in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, 2021, p. 679-689. doi: 10.18653/v1/2021.eacl-main.56.
24. A. Khalil, M. Jarrah, M. Aldwairi, et M. Jaradat, « AFND: Arabic fake news dataset for the detection and classification of articles credibility », *Data Brief*, vol. 42, p. 108141, juin 2022, doi: 10.1016/j.dib.2022.108141.
25. A. R. Mahlous et A. Al-Laith, « Fake News Detection in Arabic Tweets during the COVID-19 Pandemic », *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no 6, 2021, doi: 10.14569/IJACSA.2021.0120691.
26. H. Himdi, G. Weir, F. Assiri, et H. Al-Barhamtoshy, « Arabic Fake News Detection Based on Textual Analysis », *Arab. J. Sci. Eng.*, vol. 47, no 8, p. 10453-10469, août 2022, doi: 10.1007/s13369-021-06449-y.