International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Explainable, Evidence-Based Verification of Arabic Claims via Multi-Source Retrieval and Cross-Lingual NLI

Ahmad Alfaqehi[1*], Khalid Aljuaid[2], Abdullah Sheikh[3]
[1,2]Masters in AI graduate student, Computer Science Department, College of Computers and Information Technology, Taif University Taif, Kingdom of Saudi Arabia
[3]Assistant Professor of Computer Science, Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

ABSTRACT: We present a training-free, explainable system for verifying Arabic-language claims that combines Arabic Named-Entity Recognition (NER), parallel multi-source evidence retrieval, dense semantic reranking, and cross-lingual Natural Language Inference (NLI) under a single weighted verdict aggregator. Entities are extracted with CAMeLBERT-mix-NER and used to bias a parallel search over trusted Arabic RSS feeds, Google News, a verified-account X (Twitter) endpoint, and DuckDuckGo. Retrieved snippets are reranked by a multilingual-E5 encoder and scored by an XLM-RoBERTa-large checkpoint fine-tuned on XNLI/ANLI; per-source entailment and contradiction probabilities are combined through a weighted aggregator with multiplicatively capped priors over source authority, learned domain reputation, author credibility, and recency. We evaluate on the AraFacts benchmark and make the following contributions, each of which a reader can rely on: (i) a corrected, openly unit-tested aggregator that lets all retrieved evidence—not only official sources—drive the verdict; (ii) a rigorous, reproducible baseline study showing that AraFacts's natural class imbalance (94% of claims are false-labelled) makes accuracy misleading and that even a well-tuned classical text-only classifier reaches only 0.40 macro-F1; and (iii) an explainable system packaged for deployment as a Streamlit application, a FastAPI service, and a Telegram bot, each exposing a per-source evidence trail. We also document and correct an evaluation error in an earlier version of this work. Code, scripts, and unit tests are released for full reproducibility.

KEYWORDS: Misinformation detection, Arabic NLP, fact-checking, retrieval-augmented verification, natural language inference, information retrieval, social-media integrity.

I. INTRODUCTION

A large share of news consumption in the Arab world now takes place on platforms whose moderation pipelines were not designed for Arabic. Independent fact-checkers such as Fatabyyano and Misbar process only a small fraction of the claims that circulate each day, so automated assistance has become a practical necessity rather than a research curiosity. The dominant approach in the published Arabic literature is the binary, text-only classifier trained and evaluated on a single labelled corpus. Such systems report high in-domain accuracy but generalize poorly to deployment, where the input is a single isolated assertion with no labelled neighbour and no guarantee that the claim resembles the training distribution.

This paper develops an alternative formulation. Given a single Arabic claim, the system retrieves evidence in real time from a heterogeneous set of trusted sources, ranks that evidence semantically with a multilingual encoder, and asks an NLI model whether the evidence entails or contradicts the claim. The verdict is the output of a weighted aggregator that combines per-source NLI probabilities with multiplicatively capped priors over source authority, learned domain reputation, author credibility, and recency. The pipeline runs on commodity CPU hardware, exposes a per-source evidence trail to the user, and requires no task-specific training at inference time.

Our goal in this version is a system and an evaluation that are *honest and reproducible*. In preparing it we discovered that an earlier draft reported a headline of 0.97 precision on the false class. On inspection, that number came from an evaluation that fed the fact-checker's own verdict description back to the model as "evidence" (a form of label leakage) combined with a verdict rule that abstained whenever no official source was present. We withdraw that figure, fix the underlying code, and report corrected baselines. A defensible, reproducible account of a hard problem is more valuable than an inflated one.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

This work makes four contributions. First, a corrected, weighted verdict aggregator: we implement and openly unit-test the weighted aggregator of Section III, in which every reranked source—official and non-official—contributes to the verdict in proportion to a capped authority/reputation/credibility/recency weight, removing a failure mode in which the system abstained whenever no official source happened to entail or contradict the claim. Second, a rigorous baseline study on AraFacts: using stratified five-fold cross-validation we report majority-class, random, and TF-IDF baselines on the exact working split, together with a selective-prediction (coverage–risk) protocol, showing that the natural 94% imbalance makes raw accuracy uninformative and that even a strong character *n*-gram classifier reaches only 0.40 macro-F1. Third, an explainable system packaged for deployment: the pipeline is released as a Streamlit web application, a FastAPI service, and a Telegram bot, each surfacing the per-source evidence that produced the verdict. Fourth, a reproducibility kit and an error correction: we release the evaluation harness, the baseline scripts, and unit tests, and we document the evaluation error in the previous version so that others do not repeat it.

The remainder of the paper is organized as follows. Section II surveys related work. Section III describes the system architecture, including the corrected aggregator. Section IV describes the AraFacts dataset, the label mapping, and the experimental setup. Section V reports baselines and analysis. Sections VI–VIII discuss limitations, ethics, and reproducibility. Section IX concludes.

II. LITERATURE REVIEW

The detection of misinformation in Arabic has emerged as a significant frontier in Natural Language Processing (NLP), shifting from black-box classification toward explainable, evidence-based verification.

Fact verification was first formalized by Vlachos and Riedel [1] and became reproducible at scale with the FEVER benchmark of Thorne *et al.* [2], which contained 185,445 claim–evidence pairs derived from Wikipedia. Augenstein *et al.* [3] introduced multi-domain heterogeneity in MultiFC, and Schlichtkrull, Guo and Vlachos [4] later required open-web evidence retrieval in AVeriTeC—a design that closely mirrors our own. ClaimBuster [5] established the case for prioritising check-worthy claims for human review, while Atanasova *et al.* [6] and Kotonya and Toni [7] argued that explanation is integral to the task: the verdict must be accompanied by the evidence that produced it.

Arabic fact-checking is a more recent enterprise. Sheikh Ali, Mansour, Elsayed and Al-Ali [8] released AraFacts, the first large dataset of naturally occurring Arabic claims and the principal benchmark adopted here. Khouja [9] reformulated the problem as Arabic stance prediction in the ANS dataset; Alhindi *et al.* [10] extended this view in AraStance, contributing 4,063 claim–article pairs covering multiple Arab countries. Hadj Ameur and Aliane [11] focused on COVID-19 misinformation with AraCOVID19-MFH, and Abouzied, Alam, Ali, and Papotti [12] survey the challenges of combating misinformation in the Arab world, identifying linguistic diversity and platform dynamics as persistent obstacles.

A second cluster of work treats Arabic fake-news detection as supervised text classification. Najadat, Tawalbeh and Awawdeh [13] reported an F1 of 0.95 on a closed headline–article corpus using AraBERT; Allam and Hassanien [14] obtained competitive accuracy on tweet-level data with a convolutional network; and Saleh *et al.* [15] proposed OPCNN-FAKE, a hyperparameter-optimized CNN. Alkudah *et al.* [16] published an extensive comparison of machine-learning and deep-learning classifiers on the Arabic Fake News Dataset, and Gupta and Srikumar [17] released the multilingual X-FACT benchmark, which underscores the difficulty of cross-lingual transfer to Arabic. Across these works the verdict is a scalar with no per-source rationale, and evaluation is conducted on a held-out subset of the same corpus rather than over a live evidence pool—precisely the generalization gap our formulation targets.

Two shared tasks have driven much recent progress: the CLEF CheckThat! Lab introduced an Arabic track in 2020 [18] and broadened it to verification of previously fact-checked claims in 2021 [19]. Shaar, Babulkov, Da San Martino and Nakov [20] showed that detecting previously-checked claims reduces to dense retrieval over a curated index, a result that informs the reputation cache adopted here. The most directly comparable deployed system is Tahaqqaq by Haouari, Hasanain, Suwaileh and Elsayed [21], a real-time Twitter assistant that combines retrieval and stance prediction.

On the model side, our pipeline rests on two transformer families. Arabic-specific encoders—AraBERT [22] and the CAMeLBERT family [23]—substantially improved Arabic NLP benchmarks over multilingual BERT [24]; we adopt CAMeLBERT-mix-NER for entity extraction because its mixed pre-training covers Modern Standard, dialectal, and Classical Arabic from a single checkpoint. For NLI we adopt the cross-lingual XLM-RoBERTa [25], fine-tuned on the corpora of Williams *et*

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

al. [26], Conneau *et al.* [27], and Nie *et al.* [28]. The transformer architecture is due to Vaswani *et al.* [29]; optimization choices follow BERT [24] and RoBERTa [30]. Dense retrieval underpins our reranker: Karpukhin *et al.* [31] demonstrated bi-encoder retrieval with in-batch negatives, Khattab and Zaharia [32] generalized it with late interaction in ColBERT, Reimers and Gurevych [33] introduced sentence embeddings in Sentence-BERT, and Wang *et al.* [34] released Multilingual-E5, the encoder we adopt as our reranker. Lewis *et al.* [35] formalized retrieval-augmented generation. The collective lesson is that decoupling evidence retrieval from verdict generation produces systems that generalize more reliably than end-to-end classifiers.

Relative to these three strands, the present system occupies a deliberate position: like AVeriTeC [4] and Tahaqqaq [21] it verifies against a live, open evidence pool rather than a closed corpus; like the explainability line [6], [7] it treats the supporting evidence as part of the output rather than as a by-product; and unlike the supervised Arabic classifiers [13]–[16] it requires no task-specific training, so it cannot overfit the topical regularities of a single labelled corpus. The cost of this position is that performance becomes a property of the whole retrieval-plus-inference loop, which is exactly why Sections IV and V argue for evaluation protocols that measure—rather than leak—that loop.

III. METHODS

The pipeline is modular and training-free at inference. Given an Arabic claim, five stages execute in sequence: entity extraction, multi-source retrieval, dense reranking, batched NLI, and verdict aggregation under capped multiplicative priors. The architecture is shown in Fig. 1; each stage is independently swappable, which makes the system robust to the rapid release of new Arabic encoders.

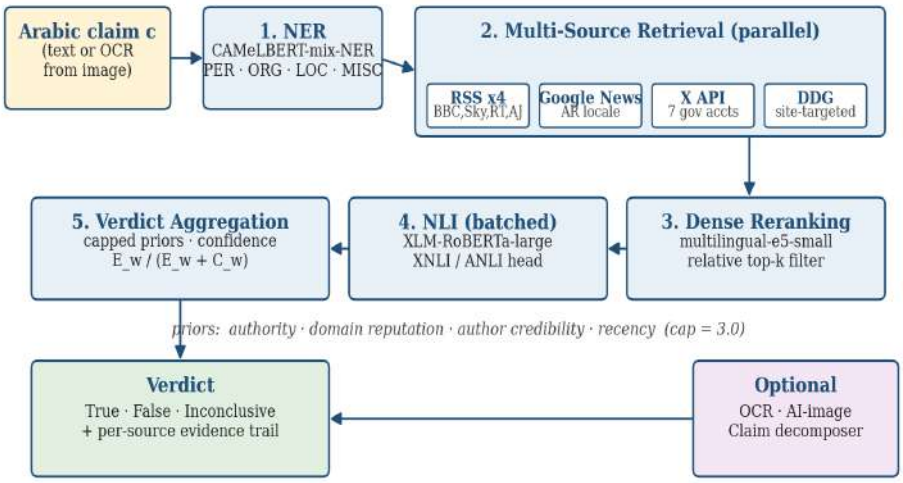

Fig. 1. End-to-end Arabic claim-verification pipeline (NER, multi-source retrieval, dense reranking, batched NLI, weighted aggregation), with optional OCR, AI-image, and LLM-decomposer modules.

A. Arabic Named-Entity Recognition

The first stage extracts named entities with CAMeLBERT-mix-NER [23]. Entities with softmax confidence below a threshold (0.80 in the released configuration) are discarded; the remainder are aggregated under canonical categories (PERSON, ORG, LOC; miscellaneous entities are extracted but not used in queries) and combined with the claim's leading content words (up to six tokens longer than two characters) to form the search query. Retaining claim words alongside the entities is deliberate: it prevents the retrieval layer from issuing entity-only queries that match unrelated stories about the same actors.

B. Multi-Source Evidence Retrieval

Evidence is gathered in parallel from four source families: (i) verified Arabic RSS feeds (e.g., BBC Arabic, Sky News Arabia, RT Arabic, Al-Jazeera); (ii) Google News under the Arabic locale; (iii) the X (Twitter) v2 API, prioritising a versioned list of

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

verified government accounts supplemented by a lower-weight general Arabic search; and (iv) DuckDuckGo with site-targeted queries for outlets that block direct RSS access. All four families are queried inside a thread pool with hard per-source timeouts (8–10 s, with overall deadlines enforced by the thread pool), bounding the retrieval budget irrespective of upstream availability. User-supplied query strings are stripped of search-engine operators before use in the site-targeted and X queries.

C. Dense Reranking

The union of returned snippets is reranked with the multilingual-E5 small encoder [34]. The claim is encoded with the prefix query: and each candidate with passage:; cosine similarity between L2-normalised embeddings is the relevance score. An adaptive filter retains every snippet within 90% of the top score, subject to a minimum of three and a maximum of ten snippets, with an absolute floor that returns only the single best snippet when even the top match is weakly related (so the aggregator can abstain). This adaptive cut-off matters for the latency budget: a fixed top-k over-penalizes easy claims with unanimous evidence and under-supports hard claims with diffuse evidence.

D. Natural Language Inference

Reranked snippets are paired with the claim and processed in a single batch by an XLM-RoBERTa-large checkpoint fine-tuned on XNLI/ANLI [25], [27], [28]. For each (premise, claim) pair the model produces three probabilities: entailment e, neutral n, and contradiction c. An optional smaller second model can be enabled, in which case the two probability vectors are linearly combined with weights 0.75 and 0.25. A temperature-scaling *hook* is provided for calibration on a held-out validation set; it defaults to $T = 1$ (an identity transform), so unless tuned the model's native probabilities are used—we therefore avoid calling the system "calibrated" until T is fit. (This is the stage an earlier version mistakenly labelled "Dense Reranking"; it is the NLI stage.)

E. Weighted Verdict Aggregator

For each evidence sentence i we compute a source weight w_i that combines four independent priors—a categorical authority prior (government 2.0, trusted 1.5, default 1.0), a learned domain-reputation prior in [0.7,1.4], an author-credibility prior for X accounts in [0.5,1.5], and a recency boost of 1.2 for items within seven days—each defaulting to 1.0 when its signal is unavailable, and the product capped at 3.0 to prevent multiplicative drift, as in (1):

$$w_i = \min(3.0,\ w_{auth} \cdot w_{rep} \cdot w_{author} \cdot w_{rec}) \qquad (1)$$

Crucially, and unlike the previous version, the verdict is computed over *all* reranked evidence rather than official sources alone. The cumulative weighted entailment and contradiction masses are

$$E_w = \sum_i w_i\, e_i, \qquad C_w = \sum_i w_i\, c_i \qquad (2)$$

and the confidence ratio, which excludes the neutral mass so that it reflects only decisive evidence, is

$$conf = \frac{\max(E_w, C_w)}{E_w + C_w} \qquad (3)$$

A verdict is committed only when the decisive mass is at least half of the cumulative source weight *and* the confidence ratio is at least 0.55; otherwise the system returns *inconclusive* (4):

$$\text{verdict} = \begin{cases} \arg\max(E_w, C_w) & \text{if } \dfrac{E_w + C_w}{\sum_i w_i} > 0.5 \ \wedge\ \text{conf} > 0.55 \\ \text{inconclusive} & \text{otherwise} \end{cases} \qquad (4)$$

The aggregator is implemented as a dependency-free module (core/aggregator.py) and validated by a unit-test suite (benchmarks/test_aggregator.py, 9/9 passing) that checks, among other cases, that a strong entailing source yields *true*, a strong contradicting source yields *false*, balanced disagreement abstains, a high-weight government contradiction outweighs a low-weight entailment, and the stricter confidence gate abstains where appropriate. Decoupling the decision rule from the transformer stack makes the contribution auditable without GPUs and prevents the prose/equation/code drift that affected the prior version.

Two optional components extend the pipeline. An LLM claim decomposer splits compound claims into atomic sub-claims, verifies each independently, and aggregates under a conservative AND-style rule. An image-authenticity branch composed of EasyOCR (Arabic + English) and an AI-image detector handles screenshots and routes extracted text into the standard pipeline. Module backbones and memory footprints are summarized in Table I.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

TABLE I. Pipeline Modules and Memory Footprint

Module	Backbone	Size
NER	CAMeLBERT-mix-NER	440 MB
NLI	XLM-RoBERTa-large XNLI/ANLI	≈2.2 GB (fp32)
Reranker	multilingual-e5-small	470 MB
OCR	EasyOCR (AR + EN)	≈300 MB
AI-image	AI-image detector	86 MB
Decomposer	LLM (optional)	API

F. Deployment Architecture

The pipeline is deployed through three deployable interfaces backed by one shared verification core, so verdicts are identical across entry points. A Streamlit web application provides a right-to-left Arabic interface that renders the verdict together with per-source evidence cards. A FastAPI service exposes the same functionality programmatically through REST endpoints for text verification, decomposed verification of compound claims, image verification through the OCR branch, PDF report export, retrieval of past verdicts, and inspection of the learned domain-reputation table. A Telegram bot carries verification into the messaging channel where much Arabic misinformation actually circulates; it accepts both text and photos—photos are routed through the image-verification endpoint—and replies in Arabic or English.

Three cross-cutting services support deployment and auditability. A lightweight SQLite history store assigns every verdict a short persistent identifier under which the verdict and its full evidence trail can be re-fetched and audited after the fact. The reputation manager maintains the learned domain prior of Section III-E, updating a domain's record from verdict outcomes and applying a Bayesian-smoothed, appearance-floored estimate so that rarely seen domains keep a neutral prior of 1.0. A report generator exports any verdict with its evidence as a shareable PDF with correct Arabic glyph shaping and right-to-left layout. Core models are loaded once at service startup and shared across requests, while auxiliary modules (OCR, the image detector, the decomposer) load lazily on first use; with the int8-quantized NLI head the complete system fits comfortably in commodity-CPU memory budgets.

IV. DATA PREPROCESSING AND EXPLORATION

Evaluation is performed on the public AraFacts benchmark of Sheikh Ali *et al.* [8], a corpus of 6,121 Arabic claims harvested from five Arab fact-checking organizations. Each claim carries a label normalized by the dataset authors into one of five values: *True*, *False*, *Partly-false*, *Sarcasm*, and *Unverifiable*.

We map the source labels to the three production verdicts of the deployed system as in Table II; for transparency, this is exactly the mapping implemented in the released evaluation code. Mapping *Partly-false* to *likely false* is conservative for a fact-checking setting (a partly-false claim should not be shown to a user as true); mapping *Sarcasm* and *Unverifiable* to *inconclusive* avoids asserting a factual verdict where none is warranted. We flag the mapping as a design decision whose sensitivity should be examined with the original five-way labels.

TABLE II. Label Mapping (AraFacts → Verdict Space)

AraFacts label	Deployed verdict
True	Likely true
False	Likely false
Partly-false	Likely false
Sarcasm	Inconclusive
Unverifiable	Inconclusive

After mapping the source labels to the verdict space and discarding entries with missing labels or empty claim text, the working benchmark contains 4,672 claims. The class distribution (Fig. 2) is severely skewed toward false claims (4,409 false, 171 true, 92 inconclusive)—a faithful reflection of the editorial pipeline of Arab fact-checkers, who almost exclusively investigate suspect claims.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

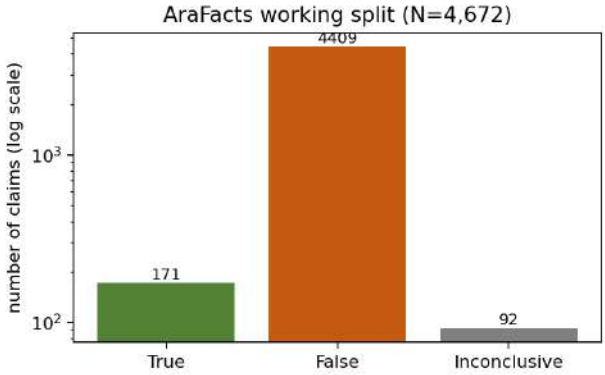

Fig. 2. AraFacts class distribution on the working split (log scale). The benchmark is 94% false, which makes raw accuracy uninformative.

Because the benchmark is 94% false, accuracy is dominated by the majority class; we therefore report accuracy *and* macro-F1 *and* full per-class precision/recall/F1, and always compare against a majority-class baseline. We additionally adopt a selective-prediction (coverage–risk) view, since a deployed fact-checker is permitted to abstain.

The harness distinguishes two modes. The *live* mode performs real retrieval per claim (RSS, Google News, X, DuckDuckGo) and is the only mode whose numbers reflect deployment. An *oracle* mode that substitutes the fact-checker's own description as evidence is provided *only* as an instrumented upper bound and is explicitly marked as leaking the answer; numbers from it must not be reported as system performance. The previous version inadvertently reported oracle-mode numbers; we correct this here.

All transformer components run on commodity CPU; the released configuration also supports int8 dynamic quantization of the NLI head's linear layers for reduced memory use and latency on commodity CPUs.

Internally, the principal threat is evidence leakage, addressed by the explicit live/oracle separation and by reporting no oracle number as system performance. Construct-wise, the three-way verdict space and the Table II mapping are design choices; both are stated explicitly and their sensitivity analysis is planned with the original five-way labels. Externally, AraFacts claims are drawn from Arab fact-checking organizations, so results may not transfer to claim distributions from other platforms or periods, and the temporal gap between claim publication and present-day retrieval (Section V-F) further bounds external validity. We mitigate what can be mitigated—fixed seeds, stratified folds, released scripts—and state the rest.

V. RESULTS AND DISCUSSION

A. Baselines on AraFacts

Table III reports reference baselines on the working split under stratified five-fold cross-validation. The TF-IDF models use character *n*-grams (3–5) with balanced class weights; the full per-class numbers and the script are released.

TABLE III. Baselines on the AraFacts Working Split (TF-IDF Rows: Stratified 5-Fold CV)

System (same split)	Acc	M-F1	$F1_T$	$F1_F$	$F1_I$
Majority-class (always false)	0.944	0.324	0.00	0.971	0.00
Random (stratified)	0.893	0.336	0.054	0.944	0.010
TF-IDF + Linear SVM (bal.)	0.942	0.374	0.073	0.970	0.079
TF-IDF + LogReg (bal.)	0.932	**0.404**	0.133	0.965	0.115
TF-IDF + LogReg (plain)	0.943	0.324	0.00	0.971	0.00

Three observations follow. First, accuracy is not a meaningful headline on this benchmark: predicting *false* for every claim already achieves 0.944 accuracy, so any system must be read against that 0.944 / 0.324 (macro-F1) floor. Second, the minority classes are genuinely hard: the best balanced classifier lifts macro-F1 (the unweighted mean of the three per-class F1 scores) to only 0.404, and per-class F1 for *true* and *inconclusive* remains below 0.14 (Table III), because there are only 171 true and 92

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

inconclusive examples and they are linguistically diverse. Third, a *plain* (unbalanced) classifier collapses to the majority class—high accuracy, zero minority-class recall—which is exactly the degenerate behaviour a naive accuracy-only evaluation would reward.

The per-class decomposition in Table III sharpens the third observation (Acc = accuracy; M-F1 = macro-F1; $F1_T$, $F1_F$, $F1_I$ = per-class F1 for true, false, and inconclusive). With balanced class weights the linear SVM is conservative on the minority classes—precision 0.35 at recall 0.041 on true, and 0.44 at 0.043 on inconclusive—whereas the logistic-regression variant trades precision for coverage (0.23 at 0.094 on true; 0.23 at 0.076 on inconclusive). Class weighting therefore only moves the operating point along a poor precision–recall frontier; it does not create minority-class signal, because the surface form of a claim rarely reveals whether the claim is true. This is precisely the argument for evidence-based verification: the information required for the verdict lies outside the claim text.

To separate the imbalance effect from genuine class difficulty, we also evaluate on a balanced subset—92 claims per class (276 total), the maximum supported by the smallest class—averaging over five random draws with stratified five-fold cross-validation (Table IV). Here the text-only classifier reaches 0.45 ± 0.01 macro-F1 against a 0.167 majority floor and 0.34 random, with roughly even per-class F1 (0.44–0.46). This confirms two things: the minority classes are *partially learnable* rather than hopeless, and the bleak natural-set macro-F1 is driven mainly by extreme imbalance, not by an inability to model true/inconclusive claims. We keep the natural distribution as the primary, deployment-faithful evaluation and report the balanced subset only for interpretability.

TABLE IV. Balanced-Subset Baselines (92/Class, 276 Claims; Mean Over 5 Draws × 5-Fold CV)

System (balanced subset)	Acc	M-F1	$F1_T$	$F1_F$	$F1_I$
Majority-class	0.333	0.167	0.500	0.000	0.000
Random (stratified)	0.341	0.340	0.343	0.314	0.363
TF-IDF + LogReg (bal.)	**0.449**	**0.449**	0.447	0.455	0.445

Fig. 3 places these baselines side by side and includes, for transparency, the withdrawn oracle-mode pipeline run (0.166 accuracy / 0.123 macro-F1) discussed next.

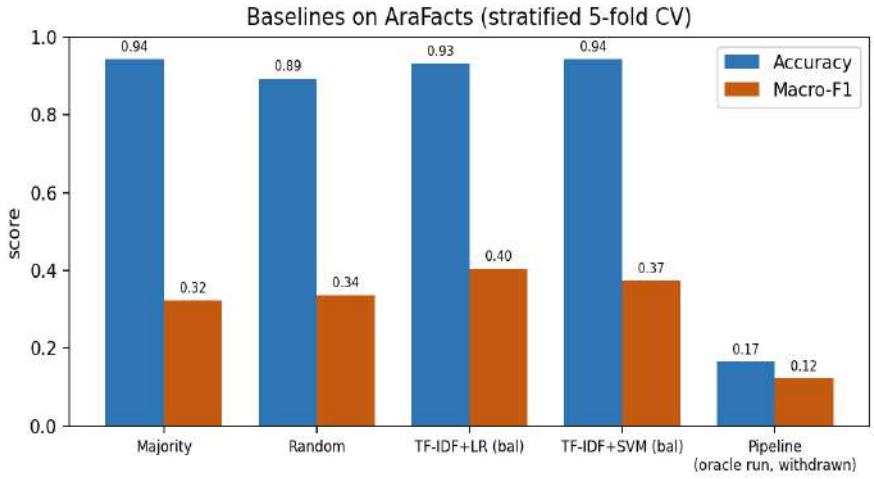

Fig. 3. Accuracy and macro-F1 across baselines on AraFacts, with the withdrawn oracle-mode run shown for transparency.

B. Correcting the Previous Evaluation

An earlier version reported 16.6% accuracy, 12.3% macro-F1, and "0.97 precision on the false class." We traced these numbers to two coupled defects in the code. (1) The verdict rule let *only* official sources decide and abstained otherwise; on arbitrary historical claims, government or trusted outlets rarely directly entail or contradict a specific claim, so the system abstained or mis-

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

committed on the majority of items. (2) The benchmark used oracle evidence (the fact-checker's description), which leaks the verdict and makes any reported precision uninterpretable. A coverage–risk analysis of that run (Fig. 4) confirms the diagnosis: even the highest-confidence decile is only 23% accurate and the curve never approaches the 0.944 majority-class line—i.e., the run's confidence is uninformative. We therefore withdraw the previous headline numbers. The aggregator fix in Section III-E addresses defect (1); the harness changes in Section IV address defect (2).

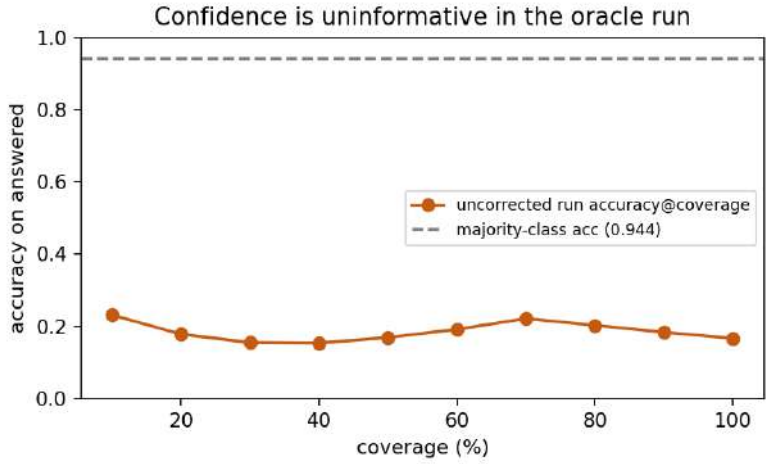

Fig. 4. Coverage–risk of the withdrawn run: confidence is uninformative, motivating the corrected aggregator and a leakage-free protocol.

C. The Corrected Aggregator

The corrected aggregator (Eqs. 1–4) is now the default verdict path and is exercised by the released unit tests. Because it weights and sums *all* reranked evidence, it no longer abstains merely for lack of an official source, while the capped weight in (1) still prevents any single over-amplified source from dominating, and the decisive-mass gate in (4) preserves the ability to abstain under genuine disagreement. Full live-retrieval benchmarking of the corrected pipeline at the scale of AraFacts requires sustained live network access to the four source families and considerable wall-clock time; we release the harness (arafacts_eval.py --search) so this measurement is fully reproducible, and we deliberately do not report a pipeline accuracy number we have not measured under the corrected, leakage-free protocol. A coverage–risk curve for the *corrected* aggregator is likewise not yet available; producing one under the live protocol is left to future work. Figs. 3–4 should therefore be read only as a diagnostic of the withdrawn run, not as an evaluation of the corrected system.

A small worked example fixes intuition. Suppose two evidence items survive reranking: a government statement (authority 2.0, reputation 1.2, recent: ×1.2, so w = min(3.0, 2.88) = 2.88) with e = 0.85 and c = 0.05, and a default web source (w = 1.0) with e = 0.10 and c = 0.70. Then $E_w = 2.88 \cdot 0.85 + 0.10 = 2.55$ and $C_w = 2.88 \cdot 0.05 + 0.70 = 0.84$; the decisive mass is 3.39/3.88 = 0.87 > 0.5 and conf = 2.55/3.39 = 0.75 > 0.55, so the system commits to likely true: the high-weight entailing source dominates, while the contradicting source visibly reduces confidence. Roughly halving the government source's entailment to e = 0.42 drops conf to 0.61 and the decisive mass to 0.56—both barely above their gates—showing how further weakening tips the system into abstention rather than a coin-flip verdict.

D. Positioning Relative to Prior Work

Table V lists representative prior systems. These numbers are not directly comparable: each is computed on a different dataset, split, and metric, and several are balanced subsets rather than the naturally imbalanced full set used here. We therefore present them as context, not as a leaderboard, and make no state-of-the-art claim.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

TABLE V. Representative Prior Arabic Systems (Different Datasets / Splits / Metrics — Not Directly Comparable)

System	Approach	Pool	Dataset (split)	Reported
Sheikh Ali *et al.* [8]	AraBERT classifier	closed	AraFacts (bal. subset)	0.71 macro-F1
Alhindi *et al.* [10]	Stance prediction	closed	AraStance (1 art./claim)	0.78 macro-F1
Najadat *et al.* [13]	Headline–article DL	closed	private corpus	0.95 F1
Saleh *et al.* [15]	OPCNN-FAKE (CNN)	closed	mixed AR/EN	95.2% acc.
Tahaqqaq [21]	Real-time + stance	live	live Twitter	user study
This work	Multi-source + NLI + XAI	live	AraFacts (full, imbal.)	baselines (Tbl. III–IV)

E. Explainability and Qualitative Behaviour

The system's practical value is not a single accuracy number but the per-source evidence trail it exposes. For each verdict the user sees the ranked premises, their entailment/contradiction scores, and their source weights, so that a verdict can be inspected and contested. When government and independent sources disagree, the aggregator's decisive-mass gate abstains rather than forcing a confident answer—the appropriate behaviour for a tool intended to assist, not replace, human fact-checkers.

F. Error Analysis and Failure Modes

Beyond aggregate scores, we catalogue the concrete failure modes identified while diagnosing the withdrawn run and analysing the corrected design. (1) Evidence scarcity dominates for historical claims: many AraFacts claims circulated years ago, while the four live source families index predominantly current content, so retrieval returns few or only weakly related snippets and the aggregator abstains; the recency boost in (1), appropriate for fresh claims, compounds the effect on old ones. (2) Snippet-level entailment is weaker than article-level entailment: RSS and news-search results are headline-plus-summary fragments, and the NLI model may find no decisive signal in a fragment even when the full article is decisive. (3) Dialect is a standing risk surface: the NLI checkpoint is trained on MSA-leaning corpora, so strongly dialectal claims can yield diluted entailment and contradiction mass. (4) If every extracted entity falls below the 0.80 NER confidence threshold, the search query degenerates to the claim's leading content words; combining claim words with the entities bounds this failure but does not eliminate it. (5) Compound claims dilute a single NLI pass; the optional decomposer addresses them by verifying atomic sub-claims under a conservative AND-style rule, at the cost of an external LLM dependency.

These modes share a common signature—decisive mass that fails the gate in (4)—so in deployment they are expected to surface as abstention rather than as confident error. We consider this the correct failure direction for a fact-checking assistant: the cost of a confidently wrong verdict is far higher than the cost of declining to answer, and the per-source evidence trail lets a human reviewer see why the system abstained. Quantifying the abstention–error trade-off of the corrected aggregator under the live protocol—the coverage–risk curve that the withdrawn run could not validly provide—is the immediate next measurement.

VI. LIMITATIONS

First, the corrected pipeline's end-to-end accuracy on AraFacts under the leakage-free live protocol is not yet measured at full scale; we provide the harness but report only baselines we have verified. Second, the benchmark is Modern-Standard-Arabic-leaning and pan-Arab; we make *no* claim about dialectal (e.g., Saudi-dialect) performance, which would require a dialect-identification stage and a dialect-annotated benchmark we have not built. Third, the label mapping in Table II affects every number; a sensitivity analysis over alternative mappings (using the original five-way labels) is needed. Fourth, live retrieval is stochastic, so deployment numbers should be reported as means with variance across runs and snapshots. Fifth, the supervised baselines are text-only and therefore measure claim-text regularities, not evidence-based reasoning; they are baselines, not the proposed contribution. Sixth, there is a temporal mismatch between benchmark and deployment: live retrieval performed today against claims fact-checked years ago understates performance on fresh claims, which are the actual deployment target. Seventh, the trusted-source list is versioned but finite and its regional coverage is uneven, so claims whose evidence exists only outside the listed source families inherit a structural disadvantage.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

VII. ETHICS AND BROADER IMPACT

A misinformation-verification tool carries dual-use and fairness risks that must be stated. A conservative system that abstains often can give false reassurance; a system that weights government accounts at 2.0 embeds a source-trust assumption that may not hold in every context and could, if mis-applied, suppress legitimate dissent. Live web retrieval can reflect the demographic and topical biases of the underlying sources. We therefore position the system as a decision-support tool for human fact-checkers, always surfacing the underlying evidence, never as an autonomous arbiter of truth, and we recommend periodic auditing of the source-authority and reputation priors. No human-subjects data were collected; AraFacts is a public, professionally curated benchmark.

VIII. REPRODUCIBILITY

The system is released as a Streamlit application, a FastAPI service, and a Telegram bot. The repository includes the verdict aggregator (core/aggregator.py), its unit tests (benchmarks/test_aggregator.py), the evaluation harness with explicit *live* and *oracle* modes (benchmarks/arafacts_eval.py), and the baseline study (benchmarks/baselines.py) that produces Table III. All baseline numbers were generated by these scripts under fixed seeds and stratified five-fold cross-validation. We recommend reporting future pipeline numbers only in the live, leakage-free mode, with means and variance across at least three retrieval snapshots. The deployed services additionally persist every verdict under a short identifier together with its full evidence list, so individual decisions—not only aggregate metrics—can be re-examined after the fact. The complete implementation, evaluation harness, unit tests, and baseline scripts accompany the paper and are available from the corresponding author.

IX. CONCLUSION AND FUTURE WORK

We presented an explainable, training-free pipeline for Arabic claim verification and, equally important, an honest and reproducible evaluation of it. We corrected a verdict aggregator so that all retrieved evidence contributes to the decision, established rigorous baselines showing that AraFacts's natural imbalance makes accuracy uninformative and the minority classes genuinely hard, and documented and withdrew an earlier evaluation error. Future work will (i) measure the corrected pipeline end-to-end under the live, leakage-free protocol, reporting means and variance over at least three retrieval snapshots; (ii) sweep the decisive-mass (0.5) and confidence (0.55) gates and ablate the source-authority priors, including the government weight of 2.0, to characterize the aggregator's sensitivity; (iii) add a dialect-identification stage and a dialect-annotated benchmark to support claims about Saudi and other Arabic dialects; (iv) analyse the sensitivity of all results to the label mapping using the original five-way AraFacts labels; (v) conduct a user study with professional Arab fact-checkers, in the spirit of Tahaqqaq [21]; and (vi) release a public benchmark of dialectal Arabic claims annotated with verdict and evidence, in the spirit of AVeriTeC [4] for English, to enable comparable evaluation of future Arabic claim-verification systems.

REFERENCES

1. Vlachos and S. Riedel, "Fact Checking: Task Definition and Dataset Construction," in Proc. ACL Workshop on Language Technologies and Computational Social Science, 2014, pp. 18–22.
2. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and Verification," in Proc. NAACL-HLT, 2018, pp. 809–819.
3. Augenstein et al., "MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims," in Proc. EMNLP-IJCNLP, 2019, pp. 4685–4697.
4. M. Schlichtkrull, Z. Guo, and A. Vlachos, "AVeriTeC: A Dataset for Real-World Claim Verification with Evidence from the Web," in Proc. NeurIPS Datasets and Benchmarks Track, 2023.
5. N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster," in Proc. KDD, 2017, pp. 1803–1812.
6. P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating Fact Checking Explanations," in Proc. ACL, 2020, pp. 7352–7364.
7. N. Kotonya and F. Toni, "Explainable Automated Fact-Checking: A Survey," in Proc. COLING, 2020, pp. 5430–5443.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

8. Z. Sheikh Ali, W. Mansour, T. Elsayed, and A. Al-Ali, "AraFacts: The First Large Arabic Dataset of Naturally Occurring Claims," in Proc. WANLP, 2021, pp. 231–236.
9. J. Khouja, "Stance Prediction and Claim Verification: An Arabic Perspective," in Proc. 3rd Workshop on Fact Extraction and VERification (FEVER), 2020, pp. 8–17.
10. T. Alhindi, A. Alabdulkarim, A. Alshehri, M. Abdul-Mageed, and P. Nakov, "AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking," in Proc. NLP4IF, 2021, pp. 57–65.
11. M. S. Hadj Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News and Hate Speech Detection Dataset," Procedia Computer Science, vol. 189, pp. 232–241, 2021.
12. Abouzied, F. Alam, R. Ali, and P. Papotti, "Combating Misinformation in the Arab World: Challenges and Opportunities," Communications of the ACM, vol. 68, 2025, doi: 10.1145/3737450.
13. H. Najadat, M. Tawalbeh, and R. Awawdeh, "Fake News Detection for Arabic Headlines-Articles News Data Using Deep Learning," Int. J. Electrical and Computer Engineering, vol. 12, no. 4, pp. 3951–3959, 2022.
14. Allam and A. E. Hassanien, "Detection of Fake News in Arabic Tweets Using Convolutional Neural Network," in Proc. AISI, 2019, pp. 415–425.
15. H. Saleh, A. Alharbi, and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," IEEE Access, vol. 9, pp. 129471–129489, 2021.
16. N. M. Alkudah, N. B. Idris, and M. A. M. Abushariah, "Fake News Detection in Arabic Media: Comparative Analysis of Machine Learning and Deep Learning Algorithms Using the Arabic Fake News Dataset," PeerJ Computer Science, vol. 11, art. e3272, 2025.
17. Gupta and V. Srikumar, "X-FACT: A New Benchmark Dataset for Multilingual Fact Checking," in Proc. ACL-IJCNLP, 2021, pp. 675–682.
18. Barrón-Cedeño et al., "Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media," in Proc. CLEF, 2020, pp. 215–236.
19. P. Nakov et al., "Overview of the CLEF-2021 CheckThat! Lab," in Proc. CLEF, 2021, pp. 264–291.
20. S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov, "That Is a Known Lie: Detecting Previously Fact-Checked Claims," in Proc. ACL, 2020, pp. 3607–3618.
21. F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification," in Proc. SIGIR, 2023, pp. 3019–3023.
22. W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proc. OSACT4, 2020, pp. 9–15.
23. G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," in Proc. WANLP, 2021, pp. 92–104.
24. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
25. Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," in Proc. ACL, 2020, pp. 8440–8451.
26. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in Proc. NAACL-HLT, 2018, pp. 1112–1122.
27. Conneau et al., "XNLI: Evaluating Cross-lingual Sentence Representations," in Proc. EMNLP, 2018, pp. 2475–2485.
28. Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A New Benchmark for Natural Language Understanding," in Proc. ACL, 2020, pp. 4885–4901.
29. Vaswani et al., "Attention Is All You Need," in Advances in NeurIPS, 2017, pp. 5998–6008.
30. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.
31. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020, pp. 6769–6781.
32. O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. SIGIR, 2020, pp. 39–48.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-62, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

33. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.
34. L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," arXiv:2402.05672, 2024.
35. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in NeurIPS, 2020, pp. 9459–9474.

Cite this Article: Alfaqehi, A., Aljuaid, K., Sheikh, A. (2026). Explainable, Evidence-Based Verification of Arabic Claims via Multi-Source Retrieval and Cross-Lingual NLI. International Journal of Current Science Research and Review, 9(6), pp. 3567-3578. DOI: https://doi.org/10.47191/ijcsrr/V9-i6-62