2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

Claim Verification in Persian: A Hybrid Pipeline with LLMs and Rationale-Quality Classification

Sara Bourbour Hosseinbeigi*, MohammadAli SeifKashani, Fateme Aali
*- Corresponding author: Information Technology Department, Tarbiat Modares University, Tehran, Iran
2- Computer Engineering Department, Sharif University of Technology, Tehran, Iran
3- Information Technology Department, Tarbiat Modares University, Tehran, Iran
s.bourbour@modares.ac.ir - ma.seifkashani@ce.sharif.edu - f.alee@modares.ac.ir

Abstract

This study presents a Persian claim-verification pipeline for assessing the relation between social-media posts and news articles. The proposed system combines encoder-based evidence classification, large language model verification, and rationale-quality filtering. Human-annotated post–article pairs are used to train LaBSE- and ParsBERT-based classifiers, while LLM-generated rationales are audited to detect weak or hallucinated explanations.

Class-conditional oversampling is applied to improve evidence classification under label imbalance. Experimental results on a confidential Persian corpus show that LaBSE with Class 0 duplication achieves the best evidence-classification performance, reaching 0.71 accuracy and 0.80 F1-score. The findings suggest that encoder models provide stable evidence detection, while LLMs are more useful as rationale generators when combined with quality control.

Research method

The proposed pipeline consists of five main stages. First, social-media posts and news articles are pre-linked using multilingual sentence embeddings and cosine similarity. Second, each post–article pair is classified into support, refute, or unrelated using pretrained encoder models such as LaBSE and ParsBERT. Third, low-confidence or negative cases are passed to instruction-following LLMs to generate short Persian rationales. Fourth, human annotators audit these rationales and assign a valid/invalid rationale-quality label. Finally, the verified rationale-quality signal is used for gating, re-ranking, and class-conditional oversampling to improve evidence classification and reduce unreliable explanations.

Post + Article → Semantic Pair Linking → Encoder-Based Evidence Classification → LLM Rationale Generation → Human Rationale Audit → Rationale-Quality Gating → Final Verification Decision

conclusion

This study proposed a hybrid Persian claim-verification pipeline that combines encoder-based evidence classification, LLM-based rationale generation, human rationale auditing, and rationale-quality filtering. The results show that LaBSE and ParsBERT are more reliable for evidence classification than LLM-only baselines. LLMs are useful for generating explanations, but their outputs require quality control to avoid vague, off-claim, or hallucinated rationales. Overall, the proposed pipeline improves interpretability and reliability in Persian claim verification, especially when class imbalance is handled through class-conditional duplication.

The purpose of the research

The main purpose of this research is to design a reliable Persian claim-verification pipeline that can determine whether a social-media post supports, refutes, or is unrelated to a news article. Since Persian online content contains informal writing, implicit claims, and heterogeneous news sources, simple similarity-based methods are often insufficient. This study focuses on improving verification reliability by combining evidence classification, LLM-based reasoning, human validation, and rationale-quality assessment. The goal is not only to predict stance labels, but also to provide more trustworthy and interpretable evidence for downstream fact-checking.

Practical or simulation results

Best Performing Model:

LaBSE + Duplicate Class 0
Accuracy: 0.71
F1-score: 0.80
Precision: 0.68
Recall: 0.97

Key Findings:
• Encoder-based models were more stable than LLM-only baselines.
• Class-conditional duplication improved evidence classification.
• GPT-4o achieved acceptable recall, but lower stability than LaBSE and ParsBERT.
• The best balance between precision and recall was obtained by LaBSE with Class 0 duplication.

References

[1] Shaar, S. et al. (2020). That Is a Known Lie: Detecting Previously Fact-Checked Claims. ACL.

[2] Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

[3] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.

[4] Thorne, J. et al. (2018). FEVER: A Large-scale Dataset for Fact Extraction and Verification. NAACL-HLT.

[5] Feng, F. et al. (2020). Language-agnostic BERT Sentence Embedding. ACL.

[6] Momeni, A. S. and Khosravi, H. (2025). FarExStance: Explainable Stance Detection for Farsi. COLING.