Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PAIR-Stance: A Data-Centric Framework for Persian Social Media News–Post Stance Detection with Human–LLM Annotation and Targeted Augmentation

Sara Bourbour hosseinbeygi

s.bourbour@modares.ac.ir

Tarbiat Modares University

Fateme Alee

Tarbiat Modares University

Research Article

Keywords: Persian stance detection, social media mining, misinformation analysis, human, LLM annotation, class imbalance, targeted data augmentation

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11265263/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PAIR-Stance: A Data-Centric Framework for Persian Social Media News–Post Stance Detection with Human–LLM Annotation and Targeted Augmentation

Sara Bourbour HoseinBeigi[1*] and Fateme Aali[1]

[1]Department of Information Technology, Tarbiat Modares University, Tehran, Iran.

*Corresponding author(s). E-mail(s): s.bourbour@modares.ac.ir;
Contributing authors: f.alee@modares.ac.ir;

Abstract

The rapid dissemination of information through social media has made reliable analysis of user reactions to news increasingly important. Stance detection provides a computational mechanism for identifying whether a social media post supports, contradicts, or is unrelated to a reference news item. However, Persian stance detection remains challenging because of limited reliable resources, linguistic variation in user-generated content, and severe class imbalance. In this study, we propose a data-centric framework for Persian news–post stance detection and construct a real-world dataset of 8,447 Persian news–post pairs collected from news sources and social media platforms. The instances are labeled as support, contradict, or unrelated through a hybrid annotation protocol combining large language model-based initial labeling, human verification, and final consistency review. Several modeling approaches are evaluated, including LaBSE, XLM-RoBERTa, and a Siamese LSTM, together with different imbalance-handling strategies. Among the baseline architectures, the Siamese LSTM achieves the highest macro-F1 score of 62%. Applying weighted cross-entropy substantially improves performance, producing the best overall result with a macro-F1 of 73% and a weighted F1-score of 92.67% on the test set. A complementary rationale-based model achieves an accuracy of 0.91, a macro-F1 of 0.73, and a weighted F1 of 0.90, whereas the proposition-based approach obtains a macro-F1 of 37%. Analysis of training and validation losses further indicates stable learning behavior with no clear evidence of severe overfitting. The results highlight the

importance of annotation quality and imbalance-aware learning for real-world Persian stance detection.

Keywords: Persian stance detection, social media mining, misinformation analysis, human–LLM annotation, class imbalance, targeted data augmentation

1 Introduction

The use of social media as a news source has changed how information is produced, disseminated, and consumed, mainly because online platforms enable low-cost, rapid, and user-driven diffusion of news content [1, 2]. However, the same affordances also facilitate the spread of misinformation and fake news, which can diffuse rapidly across online social networks [3]. Consequently, misinformation analysis and automated fact-checking have become important research problems in natural language processing and computational social science [1, 2].

Within this context, stance detection has emerged as an important component of automated fact-checking pipelines, because it models the semantic relationship between a claim and a related evidence text or response [4, 5]. Rather than determining the factuality of a claim in isolation, stance detection typically classifies the relationship between two texts into categories such as support, contradict, or unrelated, which makes it closely related to natural language inference [6, 7]. In the news–post setting considered in this study, the news article is treated as the reference text and the social media post is treated as the response text, as described in Section 4.

Despite substantial progress in stance detection and fact verification, most widely used resources and benchmarks have been developed for high-resource languages, especially English [4, 5, 8]. For Persian, several important resources have been introduced, including ParsiNLU for general Persian language understanding tasks, ParsFEVER for Farsi fact extraction and verification, and FarExStance for explainable Farsi stance detection [9–11]. However, these resources primarily focus on general NLU benchmarking, claim–evidence verification, or claim-centered explainable stance detection, rather than directly modeling naturally occurring pairwise interactions between full Persian news articles and user-generated social media posts. Therefore, there is still a need for Persian datasets that capture real-world news–post relationships under noisy and naturally imbalanced social media conditions.

This gap is important because stance detection in social media settings is affected by linguistic and contextual ambiguity. User-generated posts may contain informal language, implicit references, sarcasm, figurative expressions, and partial relevance to the target claim or news item, all of which make stance classification more difficult than simple lexical matching [12, 13]. In the news–post setting, the model must infer whether the post supports, contradicts, or is unrelated to the reference news article, even when the stance is expressed indirectly or only through partial contextual cues.

Another major challenge is class imbalance, which is common in real-world classification problems and can bias learning algorithms toward majority classes [14]. In

stance detection, minority classes such as contradiction or refutation are often particularly important for misinformation analysis, but they may occur less frequently than supportive or weakly related responses in naturally collected data [12, 13]. As observed in our dataset, the initial label distribution is strongly skewed toward the support class, with contradict and unrelated instances forming much smaller portions of the data, as reported in Section 4.5.

Recent advances in large language models (LLMs) have created new opportunities for scalable annotation, semantic reasoning, and text generation in NLP tasks [15, 16]. However, LLM-based annotation is not risk-free: model outputs can be sensitive to prompt design, may contain inconsistent labels, and may propagate systematic biases if used without human validation [17, 18]. For this reason, hybrid annotation protocols that combine automatic labeling with human verification are increasingly relevant for building reliable datasets, especially in low-resource settings where fully manual annotation can be costly.

Motivated by these limitations, this paper adopts a data-centric perspective for Persian stance detection. Rather than focusing solely on model architecture, we emphasize the importance of dataset construction, annotation quality, and controlled data augmentation. Specifically, we construct a real-world dataset of Persian news–post pairs, where each instance links a social media post to a corresponding news article and is labeled as support, contradict, or unrelated. The dataset is created through a multi-stage annotation pipeline that combines LLM-based labeling with human verification to ensure both scalability and reliability.

To address class imbalance, we apply targeted data augmentation for the underrepresented contradict and unrelated classes. Rather than relying only on random oversampling, which can increase overfitting by repeatedly exposing the model to identical minority-class instances [14], we apply controlled LLM-based rewriting to existing real minority-class samples. The rewriting process is constrained to preserve the original semantic content, stance label, and relationship with the corresponding news article, while increasing lexical and syntactic diversity. Such quality-control constraints are important because LLM-based augmentation may otherwise introduce semantic drift or label inconsistency [19, 20]. From a social media mining perspective, the resulting stance labels provide a structured representation of how user-generated posts align with, challenge, or diverge from news narratives circulating across online platforms. In summary, this paper makes the following contributions:

- We construct a multi-platform Persian social media dataset consisting of 8,447 real-world news–post pairs from Telegram and X, labeled as support, contradict, and unrelated.
- We design a hybrid human–LLM annotation protocol in which LLM-generated initial labels are manually verified and reviewed to improve annotation reliability.
- We apply a targeted augmentation strategy for underrepresented stance classes and use filtering and validation steps to reduce label inconsistency in generated samples.
- We evaluate transformer-based and sequence-based stance detection models under the original setting and multiple class-imbalance handling strategies, including undersampling, oversampling, and weighted loss, with weighted F1 used as the

main comparison metric and macro-F1 used as a complementary measure of class-balanced performance.
- We additionally evaluate two complementary modeling paths: a rationale-based classifier using human-validated explanations and a proposition-based approach that models smaller aligned semantic units.

2 Related Work

2.1 Stance Detection and Fact Verification

Stance detection is commonly defined as the task of identifying the position expressed by a text toward a target, claim, topic, or event [12]. In misinformation analysis and automated fact-checking, stance detection is often formulated as a text-pair classification problem, where the system determines whether a document, headline, evidence sentence, or user response supports, refutes, discusses, or is unrelated to a given claim [4, 5, 8]. This formulation is closely related to natural language inference, since both tasks require modeling semantic relations such as entailment, contradiction, and neutrality between two texts [6, 7].

Several benchmark datasets have shaped research on verification-oriented stance detection. The Fake News Challenge dataset FNC-1 defines stance as the relationship between a news headline and an article body, using labels such as agree, disagree, discuss, and unrelated [5, 8]. FEVER, in contrast, focuses on claim verification using evidence from Wikipedia and classifies claims as supported, refuted, or not enough information [4]. These benchmarks have provided important foundations for stance detection and fact verification, but they mainly represent controlled English-language settings rather than naturally occurring Persian news–post interactions.

Stance detection becomes more challenging in social media environments, where user-generated texts may include informal language, implicit references, sarcasm, figurative expressions, and partial relevance to the target claim or news item [12, 13]. In such settings, the model must go beyond surface-level lexical similarity and infer the semantic relationship between the reference content and the response. The present study follows this real-world formulation by pairing Persian news articles with user-generated social media posts and assigning one of three stance labels: support, contradict, or unrelated, as described in Section 4.

2.2 Stance Datasets in News, Social Media, and Persian

High-quality stance datasets differ in terms of language, domain, target type, annotation design, and evidence structure. In English, several resources have been developed for stance detection in news and social media settings. The WT–WT dataset contains 51,284 English tweets related to merger-and-acquisition events, annotated with stance labels by expert annotators [21]. STANDER extends stance analysis to the news domain by providing 3,291 expert-annotated English news articles for stance detection and fine-grained evidence retrieval in the same merger-and-acquisition domain [22]. These datasets provide valuable resources for studying stance in social media and news articles; however, they are limited to English and to a specific event domain.

Persian resources have also expanded in recent years, but their task formulations are different from the news–post pairing addressed in this study. ParsiNLU provides a Persian language understanding benchmark covering tasks such as textual entailment, reading comprehension, and other NLU tasks, rather than real-world news–post stance detection [9]. ParsFEVER adapts the FEVER-style fact extraction and verification setting to Farsi, focusing on claim verification with supporting evidence rather than social media reactions to news articles [10]. FarExStance introduces an explainable Farsi stance detection dataset in which each instance includes a claim, the stance of an article or social media post toward that claim, and an extractive explanation [11].

Although these resources are important for stance detection, fact verification, and Persian NLP, they do not directly address the setting considered in this paper: pairing full Persian news articles with user-generated social media posts and analyzing their stance relationship under naturally imbalanced real-world conditions. Therefore, the present study complements existing resources by focusing on Persian news–post stance detection, where each instance consists of a news article and a corresponding social media post labeled as support, contradict, or unrelated, as described in Section 4.

2.3 Annotation Strategies and Quality Control

Annotation quality is a central factor in the reliability of stance detection datasets. Because stance labels often depend on semantic interpretation, contextual knowledge, and the distinction between related and unrelated content, annotation protocols typically require clear label definitions, multiple annotation stages, and reliability assessment [12, 23]. Inter-annotator agreement measures, such as Cohen's kappa, are commonly used to quantify the consistency of human annotations beyond chance agreement [24].

Traditional annotation strategies often rely on expert annotators or multiple human annotators per instance, followed by aggregation procedures such as majority voting, adjudication, or expert review [25, 26]. In fact-checking and stance detection datasets, evidence-based annotation has also been used to reduce ambiguity by requiring annotators to identify text segments that justify the assigned label [4, 22]. These practices help improve dataset reliability, especially when stance categories are semantically close or when the input texts contain partial relevance.

More recently, large language models have been used to support annotation by generating preliminary labels, explanations, or candidate rationales for human review [27, 28]. However, LLM-based annotation should not be treated as a fully reliable substitute for human judgment, since model outputs may be sensitive to prompt design, inconsistent across similar inputs, or affected by biases learned during pretraining [17, 18]. Therefore, hybrid human–LLM annotation protocols require validation steps in which automatically generated labels are manually checked, corrected, or filtered.

The present study follows this hybrid perspective. Initial stance labels are produced using an LLM-based annotation stage, but the labels are subsequently reviewed by human annotators and checked for consistency before being used for model training. This design aims to combine the scalability of automatic annotation with the reliability of human verification, as described in Section 4.3.

2.4 Class Imbalance in Real-World Stance Data

Class imbalance is a common problem in supervised classification, especially when some labels occur much less frequently than others in naturally collected data [14, 29]. In such settings, standard learning algorithms may become biased toward majority classes and achieve high overall accuracy while performing poorly on minority classes [14, 30]. This issue is particularly important in misinformation analysis, because minority classes such as contradiction, refutation, or disagreement may carry high practical value for detecting contested or misleading information [12, 13].

Several approaches have been proposed to address class imbalance, including resampling methods, cost-sensitive learning, class-weighted loss functions, and data augmentation [14, 31, 32]. Random oversampling can increase the representation of minority classes, but it may also increase overfitting by repeatedly presenting identical samples during training [14, 31]. In contrast, class-weighted objectives and targeted augmentation aim to improve minority-class learning while preserving the original structure of the task [32, 33].

In real-world stance detection, imbalance is not only a technical artifact but also a property of naturally occurring discourse. User responses may be dominated by supportive, neutral, or weakly related reactions, while explicit contradiction or refutation appears less frequently [12, 21]. The dataset constructed in this study follows a similar pattern: the support class is dominant, while contradict and unrelated instances are underrepresented, as reported in Section 4.5. For this reason, the experimental design emphasizes macro-F1 and class-wise F1 in addition to accuracy, because these metrics better reflect performance on minority classes than accuracy alone [32, 34].

2.5 Data Augmentation and LLM-based Generation

Data augmentation is widely used in natural language processing to increase training diversity and reduce data scarcity, particularly when labeled data are limited or class distributions are imbalanced [35, 36]. Early text augmentation methods include simple label-preserving operations such as synonym replacement, random insertion, random deletion, and random swap [35]. Although such methods are easy to apply, they may be insufficient for stance detection because stance labels depend not only on surface wording but also on the semantic relationship between a target and a response.

More task-aware augmentation methods have therefore been proposed for stance detection. Li et al. formulate stance data augmentation as a conditional masked language modeling problem, where augmented sentences are generated by conditioning on both the target and the stance label [37]. This target-aware design is important because generated examples may otherwise become inconsistent with either the target or the intended stance label [37]. Similarly, contextual augmentation methods based on pretrained language models aim to generate more fluent and context-sensitive variants than rule-based perturbation methods [38, 39].

Another related direction is counterfactual data augmentation. In this approach, examples are modified so that the label-relevant factors change while irrelevant factors are preserved as much as possible [40]. Counterfactually augmented data can help models rely less on spurious correlations and more on features that are causally related

to the target label [40]. This idea is relevant to stance detection because contradiction and unrelated examples often require controlled semantic changes rather than simple paraphrasing.

Recent work has also explored the use of large language models for synthetic data generation in text classification. LLMs can generate diverse labeled examples and reduce the cost of manual data creation, but their usefulness is not uniform across tasks and depends on factors such as task subjectivity, label clarity, and the quality of generated samples [20]. For stance detection, this limitation is important because generated posts must remain semantically aligned with the news article and the intended stance label. Without validation, LLM-generated samples may introduce label inconsistency, semantic drift, or overly artificial patterns into the training data [20, 37].

To reduce these risks, augmentation should be controlled and followed by quality filtering rather than applied indiscriminately. In this study, we use targeted LLM-based augmentation only for the underrepresented contradict and unrelated classes, and the generated samples are filtered and validated before being added to the training set, as described in Section 3.6. This design is consistent with prior findings that augmentation is most useful when it improves minority-class coverage without introducing substantial label noise [14, 20, 32].

2.6 Positioning of the Present Work

The reviewed literature shows that existing stance detection resources and methods have made substantial progress in three directions: benchmark construction for fact verification, stance detection in social media and news domains, and augmentation-based strategies for addressing data scarcity and imbalance. However, the setting addressed in this study remains different from prior work in several important aspects.

First, widely used verification-oriented benchmarks such as FNC-1 and FEVER are mainly designed for English and focus on headline–article or claim–evidence relationships rather than Persian news–post interactions [4, 5, 8]. Second, Persian resources such as ParsiNLU, ParsFEVER, and FarExStance contribute to Persian NLU, fact verification, and explainable stance detection, but they do not directly focus on pairing full Persian news articles with user-generated social media posts under naturally imbalanced conditions [9–11]. Third, although class imbalance and annotation quality are widely recognized as important issues, many studies treat them as separate preprocessing or evaluation concerns rather than integrating dataset construction, hybrid annotation, imbalance-aware training, and targeted augmentation within a single data-centric pipeline [14, 32, 41].

To address these limitations, the present study proposes a Persian news–post stance detection framework that combines four connected components: construction of a real-world news–post dataset, hybrid human–LLM annotation, targeted augmentation for underrepresented stance classes, and imbalance-aware model evaluation. This positioning does not claim that each component is entirely new in isolation. Rather, the contribution lies in adapting and combining these components for the specific problem of Persian news–post stance detection under noisy, naturally imbalanced, and low-resource conditions.

Table 1 Comparison of related stance and fact-verification resources.

Dataset	Language	Main input structure	Main focus	Direct Persian news–post pairing
FNC-1 [8]	English	Headline–article body	News stance detection	No
FEVER [4]	English	Claim–evidence	Fact verification	No
WT–WT [21]	English	Target–tweet	Social media stance detection	No
STANDER [22]	English	Target–news article	News stance and evidence retrieval	No
ParsiNLU [9]	Persian	Multiple NLU task formats	Persian language understanding	No
ParsFEVER [10]	Persian	Claim–evidence	Farsi fact verification	No
FarExStance [11]	Persian	Claim–article/post with explanation	Explainable Farsi stance detection	No
Proposed dataset	Persian	News article–social media post	Real-world Persian news–post stance detection	Yes

In addition to the primary full-text classification models, this study evaluates two complementary modeling paths. The rationale-based path examines whether human-validated explanations preserve sufficient semantic information for stance classification, while the proposition-based path investigates whether decomposing news–post pairs into smaller aligned semantic units facilitates classification. These experiments are treated as complementary analyses rather than replacements for the primary full-text stance classifier.

As summarized in Table 1, existing datasets differ in language, input structure, and task formulation. The proposed dataset complements these resources by focusing specifically on Persian news–post stance detection, where each instance directly pairs a full news article with a user-generated social media post and assigns one of three labels: support, contradict, or unrelated.

3 Methodology

3.1 Problem Formulation

We formulate Persian news–post stance detection as a text-pair classification problem. Given a news article n_i and a social media post p_i, the objective is to predict a stance label $y_i \in \mathcal{Y}$, where:

$$\mathcal{Y} = \{\text{support}, \text{contradict}, \text{unrelated}\}. \tag{1}$$

The task can be viewed as a variant of natural language inference, where the news content represents the reference text and the post represents the response text. The model must infer whether the post semantically supports, contradicts, or is unrelated to the news.

Formally, given a dataset:

$$\mathcal{D} = \{(n_i, p_i, y_i)\}_{i=1}^{N}, \quad y_i \in \mathcal{Y}, \tag{2}$$

we aim to learn a parameterized function:

$$f_\theta : (n_i, p_i) \rightarrow \hat{y}_i, \tag{3}$$

where $\hat{y}_i$ denotes the predicted stance label for the i-th news–post pair.

3.2 Overview of the Proposed Framework

We propose a data-centric hybrid framework that integrates dataset construction, annotation, targeted augmentation, and imbalance-aware modeling. The framework consists of four main components:

1. **Hybrid annotation pipeline:** multi-stage labeling combining LLM-based initial annotation and human verification.
2. **Imbalance-aware data processing:** identification and mitigation of class imbalance through class weighting and augmentation.
3. **Targeted data augmentation:** generation of synthetic samples for minority classes using LLM-based methods.
4. **Deep stance classification models:** transformer-based and sequence-based models trained under original and augmented settings.

This design explicitly prioritizes data quality and distribution over purely architectural innovation. The proposed framework follows a data-centric paradigm, where performance improvements are achieved by improving data quality and distribution rather than increasing model complexity.

3.3 Input Representation

For the transformer-based models, each news–post pair is represented as a joint input sequence:

$$x_i = \text{[CLS]}\ n_i\ \text{[SEP]}\ p_i\ \text{[SEP]}, \tag{4}$$

where n_i and p_i denote the news article and the corresponding social media post, respectively. The maximum input length is set to 256 tokens; longer sequences are truncated and shorter sequences are padded. This joint representation is used for the XLM-RoBERTa and LaBSE configurations. In contrast, the Siamese LSTM receives the news article and the social media post as two independent token sequences and encodes them using a shared subnetwork before constructing pairwise comparison features.

3.4 Model Architectures

3.4.1 Transformer-based Model: XLM-R

To capture multilingual and contextual semantics, we employ XLM-RoBERTa (XLM-R) as the primary transformer-based model. Given an input sequence x_i, the encoder produces contextual token representations:

$$H_i = \text{XLMR}_\theta(x_i), \tag{5}$$

where H_i denotes the sequence of contextual embeddings. The representation corresponding to the [CLS] token is used as the aggregate representation:

$$z_i = H_i^{[\text{CLS}]}. \tag{6}$$

The final class probabilities are computed using a linear classification layer followed by a softmax function:

$$\hat{\mathbf{y}}_i = \text{softmax}(W z_i + b), \tag{7}$$

where W and b are trainable parameters of the classification layer.

3.4.2 Transformer-based Model: LaBSE

Alongside XLM-RoBERTa, we evaluate LaBSE as a multilingual semantic encoder for Persian news–post stance detection. The news article and the corresponding social media post are provided as a paired input with a maximum sequence length of 256 tokens. Longer sequences are truncated, whereas shorter sequences are padded to the required length. The representation produced by the encoder is subsequently passed through a regularization layer and a three-class classifier to predict support, contradict, or unrelated.

The overall classification structure is kept comparable to the XLM-RoBERTa configuration so that the effect of changing the underlying multilingual encoder can be examined without substantially altering the remaining components of the classification pipeline.

3.4.3 Sequence-based Model: Siamese LSTM

The Siamese architecture consists of two branches with shared parameters, allowing the news article and the social media post to be mapped into the same representation space. Unlike the transformer-based models, the two texts are processed independently while passing through the same encoding subnetwork.

A vocabulary is constructed exclusively from the training set. Tokens occurring at least twice are retained, and special tokens are used for padding and unknown words. Each input sequence is truncated or padded to a maximum length of 50 tokens. Tokens are first mapped to 300-dimensional embeddings and then processed by a single-layer LSTM with a hidden-state size of 256.

Because the two branches share their parameters, the news and post representations are obtained as:

$$h_i^n = \text{LSTM}_\phi(n_i), \qquad h_i^p = \text{LSTM}_\phi(p_i), \tag{8}$$

where the same parameter set ϕ is used for both inputs.

To explicitly model the relationship between the two representations, the final pair representation is constructed as:

$$z_i = [h_i^n; h_i^p; |h_i^n - h_i^p|; h_i^n \odot h_i^p], \tag{9}$$

where $\odot$ denotes element-wise multiplication. This representation captures the individual information contained in each text together with their absolute difference and element-wise interaction.

The resulting vector is passed through a classification network consisting of a linear layer followed by a ReLU activation, a dropout layer with a rate of 0.3, and a final linear layer that produces logits for the three stance classes.

3.4.4 Rationale-based Complementary Model

To investigate whether human-validated rationales preserve the semantic information required for stance classification, we additionally evaluate a rationale-based complementary model. For samples with a validated rationale, the rationale text is used as the sole input to a BERT-based three-class classifier, while the final stance label of the corresponding news–post pair is used as the prediction target.

This experiment is not intended to replace the main news–post stance classifier. Instead, it examines whether validated explanations can act as a compact representation of the semantic relationship between the two texts and can therefore support quality control, difficult-case analysis, and model interpretability.

3.5 Training Strategy

3.5.1 Optimization and Base Loss

The standard cross-entropy loss is used as the base training objective for the Siamese LSTM model:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \log P_\theta(y_i \mid n_i, p_i), \tag{10}$$

where y_i denotes the ground-truth stance label. In the imbalance-aware experiments, this objective is replaced by the weighted cross-entropy loss described in the following subsection.

The Siamese LSTM is trained using the Adadelta optimizer with a learning rate of 1.0, a batch size of 32, and a maximum of 20 training epochs. To reduce instability caused by excessively large gradients, gradient clipping with a maximum norm of 5 is applied. During training, model performance is monitored on the validation set, and the best-performing checkpoint is retained. The test set is kept separate from model selection and is used only for final performance reporting.

3.5.2 Weighted Cross-Entropy Loss

Due to class imbalance, we employ a weighted cross-entropy loss. For C stance classes, the loss is defined as:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \, \mathbb{I}(y_i = c) \log P_\theta(y_i = c \mid x_i), \tag{11}$$

where w_c is the weight assigned to class c, $\mathbb{I}(\cdot)$ is the indicator function, and $P_\theta(y_i = c \mid x_i)$ is the predicted probability for class c. Larger weights are assigned to underrepresented classes to reduce majority-class bias.

3.5.3 Imbalance Handling

Instead of relying only on naive resampling, we adopt a hybrid imbalance-handling strategy consisting of class-weighted loss, targeted augmentation, and imbalance-aware evaluation using weighted F1 as the main comparison metric, with macro-F1 and class-wise F1 reported as complementary measures.

3.6 Targeted Data Augmentation

To increase the representation of the minority classes, we employ controlled LLM-based rewriting rather than unrestricted generation of entirely new social media posts. Existing real examples from the contradict and unrelated classes are rewritten using WizardLM with the objective of increasing lexical and syntactic diversity while preserving the original semantic content, stance label, and relationship with the corresponding news article.

The rewriting process is explicitly constrained so that no new factual information is introduced and the stance of the original sample is not altered. Semantically critical elements, including negation, contradiction, temporal information, referenced persons, numerical information, and causal relationships, are required to remain consistent with the original sample.

The augmented instances are therefore treated as a limited complement to the real training data rather than as a replacement for naturally collected examples. This controlled strategy is intended to increase minority-class diversity while reducing the risk of label drift and semantically inconsistent synthetic samples.

3.7 Experimental Configurations

The experiments were organized into several configurations in order to examine the effects of model architecture and class-imbalance handling separately:

1. **Original data:** training the baseline models without additional imbalance handling;
2. **Weighted loss:** retaining the original data while assigning greater importance to errors on minority classes;
3. **Resampling:** evaluating both undersampling of the majority class and oversampling of minority classes;

4. **Targeted augmentation:** controlled rewriting of minority-class samples;
5. **Proposition-based modeling:** changing the unit of analysis from full news–post pairs to smaller proposition pairs.

3.8 Evaluation Metrics

We report accuracy, macro-F1, and class-wise F1 scores. Accuracy is defined as:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} (TP_c + FP_c + FN_c)}, \tag{12}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c.

For each class c, precision, recall, and F1 are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{13}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{14}$$

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{15}$$

Macro-F1 is computed as the unweighted average of class-wise F1 scores:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c. \tag{16}$$

Because the dataset is highly imbalanced, relying only on accuracy may provide an incomplete picture of model performance, especially for minority classes [14, 32, 34]. Therefore, in this study, weighted F1 is used as the main comparison metric, while macro-F1 and class-wise F1 scores are reported as complementary measures for analyzing minority-class behavior.

3.9 Proposition-based Complementary Path

As a complementary experiment, we investigate whether modeling smaller semantic units can improve stance detection for long or multi-part texts. The news article and the corresponding social media post are first decomposed into smaller proposition-level units. Each unit contains one or two semantically related sentences and is constructed, as far as possible, to express a single independent claim.

The extracted news and post propositions are subsequently aligned according to their semantic relationship. When a sufficiently reliable semantic relationship exists between two propositions, the pair is retained as a proposition-level instance. In cases where the alignment is insufficiently clear, an LLM is used to assist the relation analysis.

The resulting proposition pairs are then used to train a Siamese model based on XLM-RoBERTa. In this setting, the model receives a proposition from the news article

and a proposition from the social media post rather than the complete original texts. This experiment is designed to evaluate whether changing the unit of analysis from full texts to smaller aligned semantic units facilitates stance classification.

4 Dataset Construction and Annotation Protocol

4.1 Data Collection

In this study, we construct a real-world dataset of Persian news–post pairs for stance detection, focusing on the relationship between verified news content and user-generated social media responses. News articles are collected from credible and established Persian news sources to improve the reliability of the reference texts. Corresponding social media posts are collected from Telegram and X and are retrieved based on temporal proximity and lexical or semantic relevance to the news content.

Unlike synthetic or benchmark datasets, the collected data reflects naturally occurring interactions in social media, where posts may contain informal language, implicit references, and varying degrees of relevance to the original news. This design enables evaluation of stance detection models under noisy and unconstrained conditions.

4.2 News–Post Pair Construction

To construct meaningful input instances, each news article is paired with a set of candidate social media posts. The pairing process follows a two-stage filtering strategy. First, an initial retrieval step identifies potentially relevant posts using keyword-based and semantic similarity signals. Second, a refinement step filters out unrelated or weakly related posts to reduce noise in the dataset. Candidate posts were restricted to a temporal window ranging from seven days before to seven days after the publication date of each news article. This event-centered window was used to reduce the inclusion of posts associated with older events, delayed reposts, and temporally unrelated discussions. Semantic similarity between candidate news–post pairs was then measured using cosine similarity, with embeddings normalized before comparison.

A cosine-similarity threshold of 0.75 was used for initial pair selection. This threshold was selected through manual inspection of candidate pairs: lower thresholds introduced a larger number of weakly related examples, whereas substantially higher thresholds disproportionately retained highly similar and predominantly supportive pairs. In addition, near-duplicate news articles were filtered using a cosine-similarity threshold of 0.95 to reduce redundancy in the resulting dataset.

Each resulting instance consists of a news article, treated as the reference text, and a social media post, treated as the response text. This formulation aligns the task with a text-pair classification setting, where the goal is to determine the semantic relationship between the two texts.

As shown in Figure 1, the dataset construction process includes data collection, pair construction, text processing, data cleaning, multi-stage labeling, analysis, and augmentation. This pipeline is designed to approximate real-world information flow, where users react to news events within a limited temporal window. By combining lexical filtering with semantic similarity, the approach reduces noise while preserving

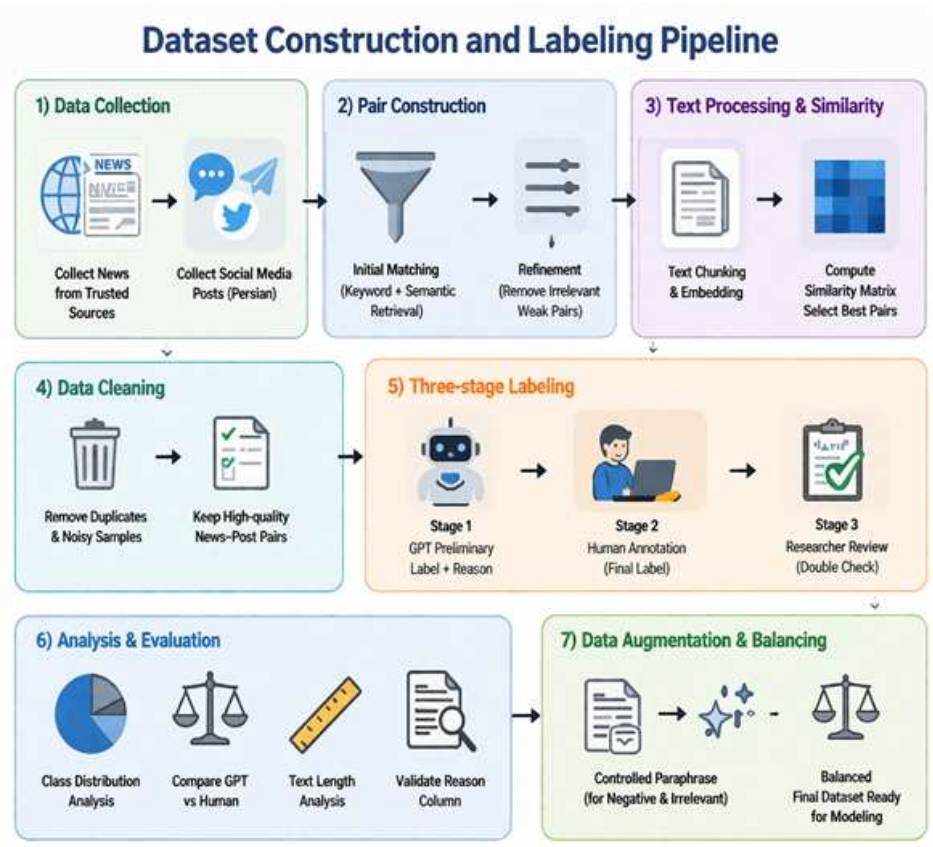


Fig. 1 Overview of the dataset construction and labeling pipeline.

diverse expressions of stance. Unlike strict one-to-one alignment, this design allows multiple posts to be associated with a single news item, reflecting realistic many-to-one relationships observed in social media.

4.3 Annotation Protocol

Given the scale and complexity of the dataset, we adopt a hybrid human–LLM annotation protocol designed to balance scalability and reliability. The annotation process consists of three sequential stages.

Stage 1: LLM-based Initial Labeling.

In the first stage, a large language model is used to assign preliminary stance labels to each news–post pair. The model evaluates the semantic relationship between the post and the corresponding news content and assigns one of three labels: support, contradict, or unrelated. This stage accelerates the labeling process by automatically handling a large portion of the dataset.

Stage 2: Human Verification.

In the second stage, human annotators review the LLM-generated labels, focusing particularly on ambiguous or borderline cases. Annotators are instructed to rely on credible interpretations of the news content and to carefully assess whether the post supports, contradicts, or is unrelated to the news.

Stage 3: Final Review and Consistency Check.

In the final stage, annotations undergo a consistency check performed by the researcher. This step aims to reduce systematic errors, resolve disagreements, and improve overall labeling quality. The multi-stage design combines the efficiency of automatic labeling with the precision of human judgment.

Agreement between LLM-generated labels and human-reviewed labels is analyzed to identify systematic discrepancies and guide dataset refinement. Samples with inconsistent or unreliable labels are removed or corrected before model training.

4.4 Label Definition

Each news–post pair is assigned one of the following three labels:

- **Support:** the post supports or confirms the information presented in the news article.
- **Contradict:** the post contradicts or challenges the information presented in the news article.
- **Unrelated:** the post does not have a meaningful semantic relationship with the news article.

This label set follows a simplified stance detection framework aligned with natural language inference formulations, where semantic relationships are commonly categorized in terms of entailment, contradiction, or neutrality.

4.5 Dataset Statistics and Class Imbalance

The final dataset consists of 8,447 news–post pairs. The label distribution is highly imbalanced: 7,227 instances belong to the support class (85.56%), 518 instances belong to the contradict class (6.13%), and 702 instances belong to the unrelated class (8.31%). This distribution shows that support is the dominant class, while contradict and unrelated form the minority classes.

This imbalance reflects naturally collected social media data, where supportive or weakly related reactions are more common than explicit contradiction. Such imbalance poses a challenge for training robust stance detection models, as standard learning algorithms may become biased toward the majority class. To address this issue, we incorporate targeted data augmentation and imbalance-aware training strategies, as described in Section 3.6.

Table 2 Distribution of stance classes in the final dataset.

Class	Number of instances	Percentage
Support	7,227	85.56%
Contradict	518	6.13%
Unrelated	702	8.31%
Total	8,447	100%

5 Experiments and Results

This section presents the experimental results of the proposed Persian news–post stance detection framework. We first describe the experimental setting and the final dataset distribution. We then evaluate the quality of the hybrid annotation process and compare the language models used for initial labeling. Subsequently, the baseline stance classification architectures are compared, followed by an analysis of different strategies for handling class imbalance. Finally, the complementary rationale-based and proposition-based modeling paths are examined, the selected model is analyzed in terms of training behavior and potential overfitting, and its performance is compared with previous work.

5.1 Experimental Setup

After data collection, preprocessing, news–post pair construction, and the hybrid annotation process, the final dataset consisted of 8,447 Persian news–post pairs. Each instance contains a reference news article and a corresponding social media post and is assigned one of three stance labels: support, contradict, or unrelated.

The final dataset is naturally imbalanced. As shown in Table 2, 7,227 instances belong to the support class, representing 85.56% of the dataset. The contradict class contains 518 instances, corresponding to 6.13%, while the unrelated class contains 702 instances, corresponding to 8.31%.

This distribution reflects the naturally occurring characteristics of news-related social media interactions, in which supportive or reinforcing reactions are considerably more frequent than explicit contradictions or completely unrelated responses. Consequently, class imbalance was treated as a central experimental challenge rather than merely as a preprocessing issue.

The data used for model development were divided into training, validation, and test sets. The training set was used to optimize model parameters, while the validation set was used to monitor performance during training and select the best-performing checkpoint. The test set was kept separate from the model-selection process and was used only for final performance evaluation.

Because of the strong class imbalance, accuracy alone was not considered sufficient for comparing the evaluated models. Macro-F1 was used to examine balanced performance across all stance classes, whereas weighted F1 was used to summarize overall

Table 3 Performance of language models for initial stance annotation.

Model	Accuracy	Precision	Recall	F1
GPT-4o	**0.8709**	**0.8843**	**0.8709**	**0.8687**
GPT-4o-mini	0.4247	0.7386	0.3095	0.4364
LLaMA	0.3498	0.7158	0.1614	0.2631

performance while accounting for the actual class distribution. These two metrics were therefore considered jointly when selecting the final model.

5.2 Annotation Quality and Initial LLM Evaluation

The dataset annotation process followed a hybrid human–LLM protocol. Large language models were used only to generate preliminary stance labels, while final labels were determined after human verification and consistency review.

Before applying the annotation pipeline to the complete dataset, several language models were evaluated to determine their suitability for producing initial stance labels. The evaluated models included GPT-4o, GPT-4o-mini, and LLaMA. Table 3 presents their performance.

As shown in Table 3, GPT-4o achieved the strongest performance among the evaluated models, with an accuracy of 0.8709 and an F1-score of 0.8687. GPT-4o-mini and LLaMA produced considerably weaker overall results, particularly in terms of recall and F1-score. GPT-4o was therefore selected as the model for generating preliminary stance labels.

Nevertheless, the GPT-4o outputs were not directly treated as final annotations. All preliminary labels were subsequently subjected to human review. Particular attention was given to difficult cases involving high lexical similarity but weak semantic correspondence, indirect references, ambiguity, or implicit stance. When the LLM-generated label disagreed with the human judgment, the human annotation was treated as the final decision. Samples for which the semantic relationship could not be determined reliably were removed or retained for additional review.

To assess the consistency of the human annotation process, Cohen's kappa was calculated after disagreement resolution. The resulting coefficient was

$$\kappa = 0.75. \tag{17}$$

This value indicates an acceptable level of agreement among the human annotators and suggests that the stance definitions and annotation guidelines provided a sufficiently consistent basis for distinguishing support, contradiction, and unrelatedness. After disagreement resolution and final review, the resulting labels were used for training and evaluating the stance detection models.

In addition to the final stance labels, short textual rationales were produced for a subset of samples, particularly difficult instances from the contradict and unrelated classes. These rationales were not automatically accepted. They were reviewed for consistency with the news article, the social media post, and the final stance label. Only

Table 4 Macro-F1 performance of the baseline architectures.

Model	Macro-F1 (%)
LaBSE	58
XLM-RoBERTa	60
Siamese LSTM	**62**

human-validated rationales were retained and subsequently used in the complementary rationale-based experiment described in Section 5.5.

5.3 The Performance of Baseline Architectures

After preparation of the final dataset, three main architectures were evaluated to compare different strategies for representing and modeling the semantic relationship between a news article and a social media post. These models were LaBSE, XLM-RoBERTa, and the Siamese LSTM.

Table 4 summarizes the macro-F1 performance of the three baseline architectures.

Among the baseline architectures, the Siamese LSTM achieved the highest macro-F1 score of 62%, compared with 60% for XLM-RoBERTa and 58% for LaBSE.

The result suggests that explicitly modeling the relationship between independently encoded news and post representations is effective for the present task. Although multilingual pretrained encoders such as LaBSE and XLM-RoBERTa are capable of extracting rich contextual and semantic information, the Siamese architecture provides an explicit mechanism for comparing the two input texts through their separate representations and pairwise interaction features.

Based on this result, the Siamese LSTM was selected as the main architecture for subsequent experiments on class-imbalance handling and performance improvement.

5.4 Effect of Class-Imbalance Handling

Because the support class represents the large majority of the final dataset, additional experiments were performed to investigate whether imbalance-handling strategies could improve model performance, particularly for the underrepresented contradict and unrelated classes.

Three strategies were evaluated using the Siamese LSTM architecture: undersampling of the majority class, oversampling of the minority classes, and weighted loss. The original Siamese LSTM without explicit imbalance handling was retained as the reference configuration.

Table 5 summarizes the results.

The baseline Siamese LSTM achieved a macro-F1 score of 62% and a weighted F1-score of 80.00%. The considerable difference between these metrics indicates that the model performed substantially better on the dominant support class than on the minority classes. Therefore, relying exclusively on the weighted metric would provide an incomplete view of model behavior.

Table 5 Effect of class-imbalance handling strategies on the Siamese LSTM.

Configuration	Macro-F1 (%)	Weighted F1 (%)
Siamese LSTM (original data)	62	80.00
Siamese LSTM + undersampling	43	80.50
Siamese LSTM + oversampling	54	91.73
Siamese LSTM + weighted loss	**73**	**92.67**

Applying undersampling reduced the macro-F1 score to 43%, while the weighted F1-score reached 80.50%. Although undersampling reduced the dominance of the support class by removing a portion of its examples, the reduction in macro-F1 suggests that discarding majority-class instances also removed useful information and diversity from the training data. Consequently, this strategy did not provide an effective improvement in overall class-balanced performance.

Oversampling resulted in a substantial increase in weighted F1, reaching 91.73%. This result indicates that exposing the model to a larger number of minority-class training instances improved its overall predictive performance. However, the corresponding macro-F1 remained at 54%, showing that the improvement across the three classes was still insufficiently balanced and that the model remained influenced by the dominant class.

The strongest results were obtained using weighted loss. Instead of changing the number of training instances, this approach increased the contribution of errors associated with underrepresented classes to the training objective. The weighted-loss configuration achieved a macro-F1 score of 73% and a weighted F1-score of 92.67%.

The simultaneous improvement of both metrics indicates that weighted learning improved not only overall predictive performance but also the model's ability to distinguish minority-class examples. Therefore, the Siamese LSTM trained with weighted loss was selected as the final model configuration for subsequent analysis.

5.5 Complementary Modeling Paths

In addition to the primary news–post stance classifier, two complementary modeling paths were evaluated. The first examined whether human-validated rationales could provide sufficient semantic information for stance classification. The second investigated whether decomposing complete news–post pairs into smaller proposition-level units could improve classification.

5.5.1 Rationale-based Model

In the rationale-based experiment, the short explanations generated for selected samples and subsequently validated by human reviewers were used as inputs to a separate classifier. Rather than processing the complete news article and social media post directly, this model attempted to predict the corresponding stance label from the validated rationale.

Table 6 Performance of the rationale-based model.

Model	Accuracy	Macro-F1	Weighted F1
Rationale-based model	0.91	0.73	0.90

The objective of this experiment was to determine whether the rationale texts preserved the essential semantic information underlying the stance relationship.

Table 6 reports the resulting performance.

The rationale-based model achieved an accuracy of 0.91, a macro-F1 score of 0.73, and a weighted F1-score of 0.90. These results indicate that the human-validated rationales contain substantial semantic information about the relationship between the news and the associated post.

However, this approach depends on the availability of a reliable rationale for each instance. It was therefore considered a complementary mechanism for analyzing difficult cases and improving interpretability rather than a replacement for the primary news–post stance classifier.

5.5.2 Proposition-based Model

The second complementary experiment investigated a proposition-based modeling strategy. In this path, the complete news article and social media post were divided into smaller semantic units so that the model could focus on more localized relationships between individual propositions.

The proposition-based configuration achieved a macro-F1 score of 37%.

This performance was substantially lower than that of the full-text Siamese LSTM. The result indicates that decomposing the complete texts into smaller units did not improve stance classification under the evaluated setting.

One possible explanation is that proposition-level segmentation removes part of the broader contextual information required for determining stance. In many news–post pairs, the relationship cannot be inferred reliably from isolated semantic units and instead depends on understanding the overall context of both texts.

Therefore, while proposition decomposition can provide a more fine-grained perspective on the input texts, the full-text news–post representation was more effective for final stance classification in the present dataset.

Overall, the complementary experiments show that validated rationales can retain useful semantic information for explanatory analysis, whereas proposition-level decomposition does not outperform the primary full-text model.

5.6 Final Model Performance and Training Behavior

After comparing the baseline architectures and the different imbalance-handling strategies, the Siamese LSTM trained using weighted loss was selected as the final model configuration.

On the test set, the selected model achieved a weighted F1-score of 92.67% and a macro-F1 score of 73%. The weighted F1-score reflects strong overall performance

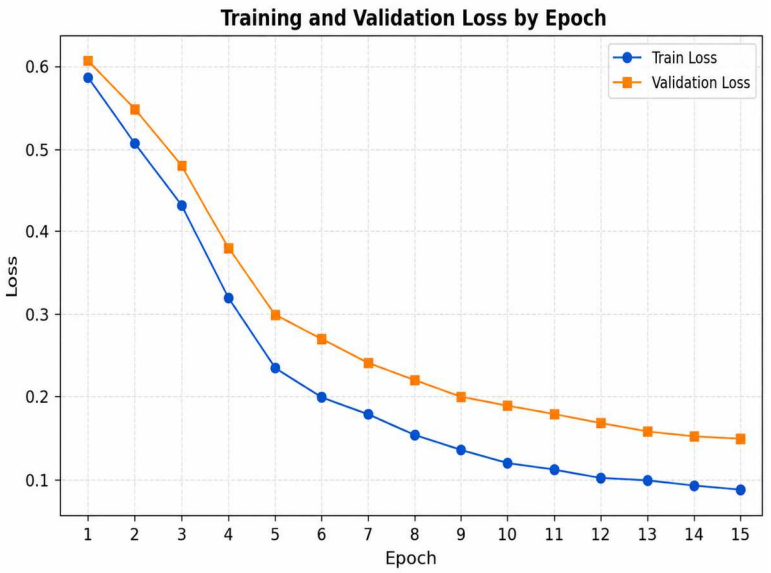


Fig. 2 Training and validation loss curves of the selected weighted-loss Siamese LSTM

under the naturally imbalanced class distribution, whereas the macro-F1 indicates that the selected configuration also provides a substantially better balance across the three stance classes than the baseline Siamese LSTM.

Although the weighted F1-score remains higher than the macro-F1 because of the dominance of the support class, the improvement of macro-F1 from 62% in the baseline configuration to 73% with weighted loss demonstrates improved performance on the minority classes.

To investigate the training behavior of the selected model and examine possible overfitting, the training and validation losses were monitored across the training epochs.

As shown in Figure 2, both the training and validation loss curves exhibit an overall decreasing trend as training progresses. The gap between the training and validation curves stays under control during the course of the observed epochs, and the validation loss does not exhibit a persistent increasing trend.

This behavior suggests that the model continued to learn useful patterns from the training data while maintaining consistent performance on unseen validation instances. Therefore, the observed loss trajectories provide no clear evidence of severe overfitting in the selected configuration.

5.7 Comparison with Previous Work

To provide a reference point for the obtained results, the final model was compared with the TESTED study, which addresses stance detection in the fake-news domain and reports a macro-F1 score of 65% [42].

The final model proposed in the present study achieves a macro-F1 score of 73%. Table 7 summarizes this comparison.

Table 7 Comparison of the final model with related work.

Study	Language	Macro-F1 (%)
TESTED	English	65
Present study	Persian	**73**

The result of the present study is higher than the macro-F1 value reported for TESTED. However, this comparison should be interpreted cautiously because the two studies differ in several important aspects, including language, dataset structure, class distribution, and evaluation protocol. As a result, it is incorrect to view the comparison as a direct, controlled benchmark between the two systems.

Nevertheless, the result indicates that the proposed approach achieves a competitive level of performance for Persian news–post stance detection despite the challenges associated with naturally imbalanced data, informal social media language, and limited Persian resources.

5.8 Discussion

The experimental results provide several observations regarding Persian news–post stance detection.

First, the comparison of the baseline architectures shows that architectural complexity alone does not guarantee better performance. Although LaBSE and XLM-RoBERTa are based on multilingual pretrained language representations, the Siamese LSTM achieved the strongest macro-F1 score among the baseline configurations. Explicitly encoding the news article and social media post separately and then modeling their relationship through comparative features therefore proved effective for the present task.

Second, the results demonstrate the importance of considering class imbalance during model training and evaluation. The baseline Siamese model obtained a macro-F1 score of 62% but a weighted F1-score of 80%, indicating that strong performance on the dominant support class can mask weaker performance on the minority classes. This observation confirms that weighted metrics should be interpreted together with macro-level measures in highly imbalanced stance datasets.

Third, the imbalance-handling experiments show that different strategies have substantially different effects. Undersampling reduced the amount of available training information and resulted in the lowest macro-F1 among the evaluated imbalance configurations. Oversampling considerably improved weighted F1, but its lower macro-F1 indicates that the improvement remained uneven across classes. In contrast, weighted loss achieved the strongest results for both macro-F1 and weighted F1 without removing or duplicating training instances. This suggests that cost-sensitive learning is particularly suitable for the naturally imbalanced distribution observed in the present dataset.

Fourth, the complementary experiments provide additional insights into the semantic structure of the task. The rationale-based model achieved a macro-F1 of 0.73

and a weighted F1 of 0.90, showing that human-validated explanations preserve a considerable amount of stance-relevant information. These rationales may therefore be useful for analyzing difficult examples and supporting interpretability. However, because the approach requires validated explanations, it cannot directly replace the primary classifier.

In contrast, the proposition-based model achieved a macro-F1 of only 37%. This result indicates that reducing the texts to smaller proposition-level units may remove important contextual information. For many news–post pairs, stance depends not only on a local semantic relationship but also on the broader meaning of the complete texts.

Finally, the analysis of the training and validation loss curves indicates stable training behavior for the selected weighted-loss Siamese LSTM. Both loss values decrease throughout training without a sustained divergence of the validation curve. Together with the test-set results, this suggests that the final model achieves improved performance without showing clear signs of severe overfitting.

Overall, the experiments demonstrate that the strongest performance in the present Persian news–post stance detection task is obtained by combining an appropriate relational architecture with an imbalance-aware training objective. The results particularly highlight the importance of weighted learning for improving the balance between overall performance and recognition of underrepresented stance classes.

6 Conclusion

In this study, we presented a data-centric framework for Persian news–post stance detection based on a real-world dataset of 8,447 news–post pairs. The framework combines hybrid human–LLM annotation, multilingual and sequence-based modeling, and class-imbalance handling to address the challenges associated with noisy social media language and highly skewed stance distributions.

Among the baseline architectures, the Siamese LSTM achieved the strongest performance, with a macro-F1 score of 62%, compared with 60% for XLM-RoBERTa and 58% for LaBSE. Subsequent experiments demonstrated that the choice of imbalance-handling strategy had a substantial effect on model performance. Undersampling resulted in a macro-F1 score of 43% and a weighted F1-score of 80.50%, while oversampling achieved a macro-F1 of 54% and a weighted F1 of 91.73%. The strongest overall performance was obtained by the Siamese LSTM trained with weighted loss, which achieved a macro-F1 score of 73% and a weighted F1-score of 92.67% on the test set. These results indicate that cost-sensitive learning provides a more effective balance between overall predictive performance and recognition of underrepresented stance classes than the evaluated resampling strategies.

The complementary experiments provided additional insights into the semantic characteristics of the task. The rationale-based model achieved an accuracy of 0.91, a macro-F1 of 0.73, and a weighted F1 of 0.90, suggesting that human-validated rationales preserve substantial stance-relevant semantic information and can support interpretability and difficult-case analysis. In contrast, the proposition-based approach achieved a macro-F1 score of 37%, indicating that decomposing complete news–post pairs into smaller semantic units may remove contextual information required for accurate stance classification.

Analysis of the training and validation loss curves of the selected weighted-loss Siamese LSTM showed stable learning behavior, with both losses decreasing throughout training and no sustained divergence of the validation loss. The observed training dynamics therefore provide no clear evidence of severe overfitting.

Overall, the findings demonstrate that reliable Persian news–post stance detection depends not only on the choice of model architecture but also on annotation quality and appropriate treatment of natural class imbalance. In particular, the results highlight the effectiveness of weighted learning for improving performance on imbalanced real-world stance data.

Future work can extend the dataset to additional Persian social media platforms and broader topical domains. Further research may also investigate evidence retrieval, reasoning-oriented architectures, and cross-lingual transfer methods to improve the handling of implicit, sarcastic, and context-dependent stance expressions.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

The dataset generated and analyzed during the current study is available from the corresponding author upon reasonable request. Requests for access may be sent to s.bourbour@modares.ac.ir.

References

[1] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017) https://doi.org/10.1145/3137597.3137600

[2] Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L.: The science of fake news. Science **359**(6380), 1094–1096 (2018) https://doi.org/10.1126/science.aao2998

[3] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018) https://doi.org/10.1126/science.aap9559

[4] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: A large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 809–819. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/N18-1074

[5] Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C.M., Gurevych, I.: A retrospective analysis of the fake news challenge stance detection task. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1859–1874. Association for Computational Linguistics, ??? (2018)

[6] Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642. Association for Computational Linguistics, ??? (2015). https://doi.org/10.18653/v1/D15-1075

[7] Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1112–1122. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/N18-1101

[8] Pomerleau, D., Rao, D.: Fake News Challenge Stage 1 (FNC-I): Stance Detection. http://www.fakenewschallenge.org/ (2017)

[9] Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., Zettlemoyer, L., Roth, D.: ParsiNLU: A suite of language understanding challenges for persian. Transactions of the Association for Computational Linguistics **9**, 1147–1162 (2021) https://doi.org/10.1162/tacl_a_00419

[10] Zarharan, M., Ghaderan, M., Pourdabiri, A., Sayedi, Z., Minaei-Bidgoli, B.: ParsFEVER: A dataset for farsi fact extraction and verification. In: Proceedings of the 10th Joint Conference on Lexical and Computational Semantics, pp. 87–93. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.starsem-1.9

[11] Zarharan, M., Hashemi, M., Behroozrazegh, M., Eetemadi, S., Pilehvar, M.T., Foster, J.: FarExStance: Explainable stance detection for farsi. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 10196–10211. Association for Computational Linguistics, ??? (2025)

[12] Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 31–41. Association for Computational Linguistics, ??? (2016). https://doi.org/10.18653/v1/S16-1003

[13] Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys **53**(5), 1–40 (2020) https://doi.org/10.1145/3395046

[14] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009) https://doi.org/10.

1109/TKDE.2008.239

[15] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)

[16] OpenAI: GPT-4 technical report (2023). Preprint at https://arxiv.org/abs/2303.08774

[17] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. Association for Computing Machinery, ??? (2021). https://doi.org/10.1145/3442188.3445922

[18] Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: Proceedings of the 38th International Conference on Machine Learning, pp. 12697–12706. PMLR, ??? (2021)

[19] Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., Li, D., Zhu, H., Cai, X., Li, Q., Liu, D., Li, T., Shen, D., Liu, T., Peng, X.: ChatAug: Leveraging ChatGPT for text data augmentation (2023). Preprint at https://arxiv.org/abs/2302.13007

[20] Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 10443–10461. Association for Computational Linguistics, ??? (2023). https://doi.org/10.18653/v1/2023.emnlp-main.647

[21] Conforti, C., Berndt, J., Pilehvar, M.T., Giannitsarou, C., Toxvaerd, F., Collier, N.: Will-they-won't-they: A very large dataset for stance detection on twitter. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1715–1724. Association for Computational Linguistics, ??? (2020). https://doi.org/10.18653/v1/2020.acl-main.157

[22] Conforti, C., Berndt, J., Pilehvar, M.T., Giannitsarou, C., Toxvaerd, F., Collier, N.: STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4086–4101. Association for Computational Linguistics, ??? (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.365

[23] Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008) https://doi.org/10.1162/coli.07-034-R2

[24] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960) https://doi.org/10.1177/001316446002000104

[25] Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics, ??? (2008)

[26] Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 614–622. Association for Computing Machinery, ??? (2008). https://doi.org/10.1145/1401890.1401965

[27] Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30), 2305016120 (2023) https://doi.org/10.1073/pnas.2305016120

[28] He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N., Chen, W.: AnnoLLM: Making large language models to be better crowdsourced annotators. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pp. 165–190. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.naacl-industry.15

[29] Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 27 (2019) https://doi.org/10.1186/s40537-019-0192-5

[30] Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5), 429–449 (2002) https://doi.org/10.3233/IDA-2002-6504

[31] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[32] Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018) https://doi.org/10.1016/j.neunet.2018.07.011

[33] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, pp. 9268–9277 (2019). https://doi.org/10.1109/CVPR.2019.00949

[34] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009) https://doi.org/10.1016/j.ipm.2009.03.002

[35] Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6382–6388. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/D19-1670

[36] Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 968–988. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.findings-acl.84

[37] Li, Y., Caragea, C.: Target-aware data augmentation for stance detection. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1850–1860. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.naacl-main.148

[38] Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 452–457. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/N18-2072

[39] Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional BERT contextual augmentation. In: Proceedings of the 19th International Conference on Computational Science, pp. 84–95. Springer, ??? (2019). https://doi.org/10.1007/978-3-030-22747-0_7

[40] Kaushik, D., Hovy, E., Lipton, Z.C.: Learning the difference that makes a difference with counterfactually-augmented data. In: Proceedings of the 8th International Conference on Learning Representations (2020). https://openreview.net/forum?id=Sklgs0NFvr

[41] Mazumder, M., Banbury, C., Yao, X., Karlas, B., Rojas, W.G., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H.R., Quaye, J., Rastogi, C., Kiela, D., Jurado, D., Kanter, D., Mosquera, R., Ciro, J., Aroyo, L., Acun, B., Chen, L., Raje, P., Bartolo, M., Eyuboglu, S., Ghorbani, A., Goodman, B., Kane, A., Shankar, S., Nushi, B., Kamar, E., Vaughan, J.W., Beutel, A., Paritosh, P., Dieng, A.B., Mattson, P., Reddi, V.J.: DataPerf: Benchmarks for data-centric ai

development (2022). Preprint at https://arxiv.org/abs/2207.10062

[42] Arakelyan, E., Arora, A., Augenstein, I.: Topic-guided sampling for data-efficient multi-domain stance detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13448–13464. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.752 . https://aclanthology.org/2023.acl-long.752/