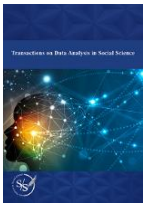
ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir



Artificial Intelligence for Sentiment Analysis of Persian Online Reviews: A Focused Review of Models, Datasets, and Evaluation

H. Ghayoumi Zadeh*

Associate Professor, Department of Electrical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.

ARTICLE INFO

Article History:
Received 20 August 2025
Received in revised form 22 October 2025
Accepted 8 December 2025
Available online 30 December 2025

Keywords:
Artificial Intelligence, Persian Sentiment Analysis, Online Reviews, Benchmark Evaluation, Domain Shift

ABSTRACT

Persian online reviews provide evidence about consumer experiences, but the usefulness of automated sentiment analysis depends on what the labels measure and where the model is tested. This focused narrative review examines benchmark construction and evaluation for Persian product and service reviews. It covers lexicon and statistical methods, convolutional and recurrent neural networks, Persian and multilingual encoders, and instruction-tuned large language models. The reviewed studies use different observation units, label definitions, and test protocols. A binary food-delivery benchmark, a three-way product recommendation dataset, and a five-level sentence polarity corpus therefore support different claims even when all report an F1 score. Comparisons within individual studies show that the size of a model advantage varies by benchmark. Evaluations across several datasets also reveal source-specific performance differences, although these differences do not identify the effect of transferring a classifier trained on one platform to another. The review proposes a reporting framework covering dataset provenance, label alignment, leakage controls, per-class metrics, and external validation. An evaluation design for transfer between platforms and a taxonomy of review errors illustrate how to apply the framework in consumer research.

1. INTRODUCTION

Online product and service reviews document how customers describe prices, delivery, quality, and support. An artificial intelligence (AI) system can classify the polarity of these comments at a scale that makes manual coding impractical. For research on consumer behavior, the output is useful only to the extent that the target is defined and the error is understood. A negative review can describe a defective product, an inconvenient delivery, or both. Even a classifier that reliably identifies negative wording does not by itself reveal which part of the customer experience needs attention.

Persian review benchmarks encode opinion in several ways. SentiPers includes manual annotation at multiple levels, DeepSentiPers extends that resource for sentence classification, and Hooshvare Lab documents food-delivery polarity and product recommendation datasets [1–3]. These resources make comparative work possible, but they

* Corresponding author: Hossein Ghayoumi Zadeh (h.ghayoumizadeh@vru.ac.ir).
Associate Professor, Department of Electrical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.

http://dx.doi.org/10.22034/tdas.2025.256843

© 2025 by the authors. Licensee T.D.A.S, Tehran, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

also pose a measurement problem. A recommendation can reflect price or expected value alongside linguistic polarity, while neutral sentiment and the absence of a recommendation need separate operational definitions.

This review asks which conclusions the available benchmarks support about AI for Persian online reviews. It examines the comparability of model results and the evidence needed to apply a classifier to a new source. Its contribution is an evidence map that connects dataset design and evaluation protocol to the claims a researcher can make about customer feedback.

2. REVIEW APPROACH

This focused narrative review used targeted searches in November 2026 through web and scholarly search, with primary records located on ACL Anthology, arXiv, publisher sites, and dataset maintainers' pages. Search terms combined Persian or Farsi with sentiment analysis and customer reviews, Digikala, SnappFood, SentiPers, or DeepSentiPers. Model-specific searches added ParsBERT, FaBERT, TookaBERT, and large language models. Dataset documentation and the citations of the retrieved papers helped identify corpus origins and later benchmark variants.

Sources were selected for direct relevance to overall sentence or document polarity in Iranian Persian product and service reviews, or for documenting a dataset used in those evaluations. Studies covering several NLP tasks were used for their review sentiment results. The quantitative comparisons cover overall polarity. An aspect-level study is included to clarify the measurement boundary, alongside methodological studies on domain adaptation, leakage, uncertainty, calibration, and documentation. The synthesis records the source, observation unit, labels, evaluation setting, and reported metrics when available. Numerical examples come from explicitly identified tables in the original studies. The evidence is purposive and no pooled effect or comprehensive estimate of model performance was calculated.

3. DATASETS AND LABELS

Table 1 summarizes the main benchmark families. Hooshvare Lab documents 70,000 SnappFood comments with equal numbers of positive and negative examples [3]. That balance makes the task suitable for controlled binary comparison. In an operational feed, the ratio of positive to negative posts can differ, and an explicit neutral option may be needed. The same documentation lists a Digikala release with 62,321 comments and three recommendation codes. The largest class is recommended, with 36,042 observations, so always predicting that class would yield about 57.8% accuracy if evaluation followed the full release distribution. A benchmark's actual test-set majority baseline should be calculated from that split.

Table 1. Persian review datasets and implications for comparison

Interpretation	Source and target	Resource
Record annotation unit and the exact variant. A binary version may remove neutral cases.	Digital-product opinions. SentiPers spans several annotation levels. DeepSentiPers supports five-level sentence polarity and derived versions.	**SentiPers and DeepSentiPers [1, 2]**
Equal class counts describe the release. Deployment prevalence requires its own estimate.	Food-delivery comments. The documented binary release contains 35,000 examples per class.	**SnappFood [3, 4]**
Keep recommendation distinct from text polarity. Verify the mapping of no idea before any alignment.	Product comments labeled recommended, not recommended, or no idea in the documented release.	**Digikala Sentiment [3]**
Name the specific corpus and test protocol. Sharing a platform does not establish equivalence with other Digikala releases.	Digikala reviews with crowdsourced positive, neutral, and negative sentiment labels.	**MirasOpinion [5, 6]**

SentiPers and DeepSentiPers form a related corpus family. A study can use five polarity levels, merge intensity levels into three classes, or remove neutral sentences and collapse intensity into two classes [1, 2, 6]. These choices change both the task and the set of observations. In addition, a review and its constituent sentences may contain overlapping text. Researchers using related variants should group source reviews and preserve augmentation ancestry when forming train and test partitions. A short dataset name is insufficient for reproducing a split or establishing independence.

Different releases from the same platform introduce another risk. MirasOpinion uses crowdsourced sentiment labels, whereas the Hooshvare Digikala release documents recommendation categories [3, 5, 6]. Treating the two as interchangeable would conceal their distinct targets. A review can be broadly positive yet decline to recommend an item because its price has risen. Likewise, a single review can praise a product and criticize delivery. The labeling rules should state how annotators handle such mixed opinions and which entity the rating concerns.

4. ARTIFICIAL INTELLIGENCE MODELS

Lexicon-based and conventional statistical models provide interpretable points of comparison. Basiri and Kabiri combined lexical and learned signals in HOMPer and compared the approach with a pure lexicon method and a Naive Bayes baseline [7]. In a modern benchmark, a majority classifier and a regularized linear model with word and character n-grams can show how much predictive signal comes from class prevalence and surface patterns. These baselines also help judge whether a more expensive model produces a meaningful improvement on the intended test set.

Convolutional and recurrent networks can learn patterns that go beyond fixed sentiment dictionaries. DeepSentiPers couples corpus augmentation with CNN and bidirectional LSTM models [2]. Shumaly et al. compared fastText and TF–IDF representations and several classifiers for Digikala data [8]. Their data construction uses a large review collection for representation learning and a pseudo-labeling procedure. The size of that collection should therefore be distinguished from the size of any independently labeled test set. In this family of studies, performance can depend on how much information the labeling and data construction process shares with the features used by the classifier.

Pretrained encoders bring contextual representations to the classification task. ParsBERT evaluates Persian-specific pretraining on sentiment and other Persian language tasks [4]. FaBERT uses Persian blogs for pretraining and reports fine-tuned results for two review sentiment corpora [6]. TookaBERT compares monolingual and multilingual encoders under a shared tuning procedure across several Persian benchmarks [9]. These studies make it possible to compare models within a reported protocol. Their results also show that an advantage on one benchmark may be much smaller on another, as Table 2 illustrates.

Generative models offer a different interface. A model can receive a description of the task and return a label without fitting a separate classifier for each benchmark. Arnob and Mahi evaluated multilingual instruction models and a Persian-adapted Gemma variant with a common prompting protocol [10]. For such comparisons, the prompt, model version, decoding settings, and parsing policy are part of the classifier. If an answer contains an explanation, a synonym, or an unexpected label, the scoring rule should state how the output is counted. Potential overlap between public reviews and model training corpora also limits claims about testing on unseen data when training provenance is unavailable.

5. WHAT THE REPORTED SCORES SUPPORT

Table 2. Selected comparisons under the protocols of the original studies

Supported interpretation	Reported scores (%)	Source and setting
In that study, the FaBERT advantage over ParsBERT was larger on DeepSentiPers.	Macro-F1, ParsBERT vs FaBERT: 74.94 vs 79.85 on DeepSentiPers and 86.73 vs 87.51 on MirasOpinion.	**FaBERT study, supervised benchmark comparison [6]**

Supported interpretation	Reported scores (%)	Source and setting
The improvement depended on the benchmark. F1 is reported as labeled in the source table.	F1, ParsBERT vs TookaBERT-Large: 80.2 vs 85.7 on DeepSentiPers and 87.6 vs 88.1 on SnappFood.	**TookaBERT study, shared encoder fine-tuning search [9]**
Source-specific performance varies under the study protocol. This comparison does not estimate supervised transfer.	Accuracy / macro-F1 for Qwen2.5-1.5B: 76.6 / 51.4 on SnappFood, 74.5 / 56.8 on Digikala, 48.0 / 41.0 on SentiPers.	**Arnob and Mahi, prompted evaluation by source [10]**

The two encoder comparisons illustrate the value of matched protocols. FaBERT exceeded ParsBERT by 4.91 percentage points on DeepSentiPers and by 0.78 points on MirasOpinion in the FaBERT study [6]. In the TookaBERT study, TookaBERT-Large exceeded ParsBERT by 5.5 points on DeepSentiPers and by 0.5 on SnappFood [9]. These differences are descriptive. The DeepSentiPers scores from the two papers should not be combined into one ranking without establishing common data versions, label mapping, test examples, and metric definitions.

The prompted comparison in [10] also illustrates why accuracy and macro-F1 deserve separate attention. Qwen2.5-1.5B achieved higher accuracy on the SnappFood subset than on the Digikala subset, whereas macro-F1 was higher on the Digikala subset. That reversal shows how the impression of relative performance can depend on the metric. It cannot isolate the contribution of platform, label distribution, text register, source sampling, or prompt behavior, since those factors were not independently controlled in the source breakdown.

Three evaluation designs answer different questions. In-domain testing estimates performance on held-out material from a corpus under a specified split. Multi-source testing reports performance for a fixed model on several sources and can reveal uneven behavior. An explicit transfer experiment trains on a declared source, fixes the checkpoint and decision rule, and evaluates on a separately defined target whose labeled examples did not guide model or prompt selection. To interpret the resulting performance gap, both domains need a shared label definition, with native-label results also reported where possible.

6. IMPLICATIONS FOR CONSUMER RESEARCH

The unit of inference should be defined before choosing the model. If the goal is to summarize opinions posted to a platform, the dataset should document which reviews were eligible, what period they cover, and how duplicate reviews were handled. If the goal is to estimate satisfaction among all customers, people who post reviews are a selected group. Automated coding can reduce annotation effort, but it does not recover the opinions of customers absent from the review sample.

Aggregate trends can also move when classification errors change. A shift in product mix, phrasing, or review length can alter the rate at which neutral or mixed comments are labeled negative. Analysts should audit a small, stratified sample with human annotation before interpreting month-to-month changes. Class-wise errors and reviewer disagreement help establish whether a trend reflects customer feedback or a change in model behavior. Reviews that contain opposing opinions about delivery and product quality warrant special attention when an overall polarity label drives a decision.

For organizations processing reviews routinely, an encoder and a prompted model should be compared on the same labeled sample, using the same label space and criteria for invalid responses. The comparison can include inference time and the percentage of cases routed to human review. An option to abstain is particularly useful for ambiguous comments. If an abstention rule is used, report coverage together with classification metrics, because performance on the retained subset can improve while the share of reviews handled automatically falls.

7. RECOMMENDED EVALUATION AND REPORTING

Define the target and publish its label map. State whether each example is a review, a sentence, or a product-specific judgment. Describe how neutral, mixed, and missing recommendations are treated. For external validation,

use a shared definition of polarity and document examples excluded during harmonization. Keep the native-label evaluation visible so readers can assess what the aligned task omits.

Protect the evaluation partition. Deduplicate before splitting, group sentences by their source review, and keep augmented examples with their parents. If the application predicts on future reviews or unseen product families, add a chronological or group-based holdout that reflects that use. In a platform transfer experiment, reserve labeled target data for final evaluation and document any access to target examples during model development.

Report accuracy, macro-F1, and class-wise precision and recall alongside the class distribution and a confusion matrix. Specify how F1 was averaged and how invalid outputs were scored. Use paired uncertainty estimates when systems are evaluated on the same test examples, accounting for clusters such as sentences from one review when relevant. Report absolute performance on both source and target, so that the size of the transfer gap can be read alongside the quality of each result.

Provide enough detail to reproduce the classifier and audit errors. For encoders, record the checkpoint, preprocessing, and split. For prompted systems, record the model version, prompt, decoding configuration, and label parser. Error analysis should sample colloquial spelling, negation, mixed opinions, and short or ambiguous comments. When review text cannot be redistributed, publish the evaluation procedure and permissible identifiers or checksums where feasible.

8. CONSTRUCT VALIDITY IN PERSIAN REVIEW MINING

A sentiment label is an operational definition. For example, a question about whether the buyer recommends a device elicits a judgment about its overall value, whereas an instruction to label sentiment asks annotators to interpret language in the review. A buyer can recommend a product while criticizing one feature. Conversely, a favorable description may accompany a decision against buying again. Before combining corpora, compare the annotation prompts and inspect a sample of disagreements between their targets. Even a plausible one-to-one mapping may remove systematic differences that matter to the social-science question.

The observation unit changes the estimand. A review containing one positive sentence about performance and two negative sentences about delivery contributes three observations in a sentence benchmark and one in a review benchmark. Counting those observations as interchangeable gives more influence to people who write longer reviews and to issues that prompt longer explanations. If a research question concerns the prevalence of dissatisfaction among reviews, the analysis should assign one observation per eligible review or explicitly define the aggregation rule. If it concerns recurring problems with delivery, the unit may instead be an aspect mention nested within the review.

Pars-ABSA illustrates the distinction between aspect and review targets. It contains annotations for individual targets within product reviews, so multiple labeled observations can come from a single source review [11]. It can guide the design of an audit for mixed comments even though its benchmark results answer a separate question. A single polarity label works for an overall summary, while an aspect label can support a decision about which feature to investigate. Test partitions and uncertainty estimates for aspect-level analysis should respect the nesting of aspect mentions within reviews.

Annotation disagreement can reveal ambiguity in the research construct. Two readers may interpret a polite complaint differently if instructions leave the neutral class undefined. During corpus design, record the label before adjudication and distinguish disagreement about tone from disagreement about the target entity. Report class-wise agreement or the distribution of disagreements along with the consensus procedure. An adjudicated label is suitable for the main benchmark, while an ambiguous subset can be used to test whether an automated system routes difficult examples to a human.

Lexical resources for Persian have long needed to handle differences in orthography and the context in which polarity-bearing words appear [12]. In online reviews, preserving negation and clause boundaries can be as consequential as normalizing character variants. A research protocol should compare raw and normalized inputs on the same test examples. Log each transformation and retain a link to the original text for error analysis. If

normalization removes repetition, punctuation, or emoji, test whether the loss disproportionately affects the detection of strongly worded complaints.

Table 3. Error categories for auditing Persian online review sentiment

What the audit records	Diagnostic example in English translation	Error category
Whether the classifier attaches negation to expectation or to product performance.	"I did not expect it to work this well."	**Negation and scope**
Whether the task asks for review polarity, product sentiment, or delivery sentiment.	"The headphones sound good, but the package arrived damaged."	**Mixed targets**
Whether category changes the evaluation of the same surface form.	"It is light" in a phone review and in a blanket review.	**Context-dependent adjective**
Whether the system detects negative evaluation without an explicit sentiment adjective.	"It only lasted until the second charge."	**Indirect dissatisfaction**
Whether annotation permits insufficient evidence and whether the model can abstain.	"Received yesterday. I have not tried it yet."	**Unresolved sentiment**

The examples in Table 3 are constructed diagnostic cases. A production audit should combine such controlled examples with naturally occurring review samples. Controlled examples help isolate the behavior of interest, while observed examples show whether the issue occurs often enough to affect the intended use. Each audit record should include the review unit, the expected label, the annotator's reason, and the system output. A comparison between two systems should use the same diagnostic cases and decision rules.

9. EVALUATION DESIGN AND QUANTITATIVE INTERPRETATION

Domain adaptation research has treated source and target product categories as separate environments because a useful feature in one category can behave differently in another [13]. The same design logic applies to Persian product and service reviews. A word associated with fast delivery may dominate a food-delivery dataset and contribute little to electronics reviews. Platform shift can also reflect changing label prevalence or a different annotation question. These changes should be measured separately where the data allow. A table of source and target prevalence, review length, product categories, and disagreement rates can identify plausible explanations for transfer errors.

Use three partitions within the source to develop a classifier: training for fitting, validation for model and threshold selection, and testing for the in-domain estimate. A target test set with independently assigned labels should be held until the model and mapping are fixed. If the target labels inform fine-tuning or prompt design, report that experiment as target adaptation and preserve a separate target test partition. The two directions should be reported separately because their source distributions, sample sizes, and class priors can differ.

Exact duplicates and near duplicates can make a benchmark look easier. In Persian reviews, short repeated phrases may occur independently, so a deduplication rule should distinguish a common phrase from repeated or copied longer text when feasible. Deduplication research has shown the value of measuring overlap in language-model training data [14]. For review classification, grouping source reviews and documenting augmentation before splitting are direct ways to protect test independence. With a pretrained model whose data are only partly documented, use a dated or private test collection when feasible and state the remaining uncertainty about pretraining exposure.

Accuracy equals the proportion of all examples correctly classified. Macro-F1 averages the class-specific F1 scores, with each class receiving equal weight. For a class, F1 is the harmonic mean of precision and recall. The scores answer different questions under imbalance. In a hypothetical test set with 60 positive, 25 negative, and 15

neutral reviews, a classifier that always predicts positive has 60% accuracy and 25% macro-F1 when classes without predicted instances receive zero F1. The example shows why an accuracy result needs a class distribution and a baseline to be interpretable.

The numerical comparisons in Table 2 can be extended with two contrasts from the same published prompted evaluation [10]. For Qwen2.5-1.5B, SnappFood exceeds SentiPers by 28.6 percentage points in accuracy and 10.4 points in macro-F1. For the Persian-adapted Gemma model, SnappFood exceeds SentiPers by 4.8 points in accuracy, while SentiPers exceeds SnappFood by 16.6 points in macro-F1. These arithmetic contrasts quantify source sensitivity under that evaluation protocol. They do not identify a causal effect of adaptation because the model families and training histories differ. They also show why selecting a metric after viewing the results can change the account of which source appears most difficult.

Compare systems with paired estimates when they predict on the same examples. For a bootstrap interval around an F1 difference, resample the same groups for both systems and recompute the difference in each replicate. The groups should match the study design, such as reviews when a benchmark includes several sentences from a review. If the paper compares many systems or slices, identify the primary comparison and treat exploratory comparisons as such. Statistical tests for NLP depend on the metric, the experimental design, and the independence assumptions [15].

Confidence and correctness should be evaluated separately. A classifier with adequate discrimination can still issue probabilities that are too high, a problem examined by Guo et al. [16]. For a review workflow that sends low-confidence cases to an annotator, choose the confidence threshold using validation data and report performance and coverage on the target test set. Check that the abstention rule does not disproportionately remove rare complaint categories. For prompted models, text such as “high confidence” is not a calibrated probability without validation.

Table 4. Proposed experiments for separating performance claims

Report	Design	Experiment
Class support, accuracy, macro-F1, per-class precision and recall, paired interval.	Fixed source train, validation, and grouped test partitions. Compare a majority baseline, a linear baseline, and an encoder.	**Within-platform reference**
Results for each direction, sample counts before and after alignment, and comparison with in-domain scores.	Train on one source and test on a held-out target with harmonized human labels. Repeat in the other direction if both sources allow it.	**Transfer between platforms**
Per-source metrics, invalid outputs, class support, and any known training overlap.	Run the same target examples through fixed prompts and checkpoint versions. Score with a documented parser.	**Prompted model comparison**
Changes in class prevalence, performance, and error types across time.	Train and tune on earlier reviews. Use an untouched later period and retain platform metadata for the audit.	**Temporal validation**
Coverage, error rates by class, and the share of negative or mixed reviews escalated.	Fix a routing threshold before the final test. Audit a sample of retained and escalated cases.	**Human review and abstention**

10. IMPLEMENTATION AND FUTURE RESEARCH

The evaluation designs in Table 4 can support a staged comparison. Begin with a small, independently annotated target set to confirm that the annotation guidelines travel between platforms. Once that condition is met, evaluate a baseline and a fine-tuned encoder on fixed partitions. Add a prompted language model to test whether it offers a meaningful gain for ambiguous comments under the same scoring rules. Model selection should weigh the

performance difference against the number of reviews requiring human review and the operating cost of a second system.

The review also identifies a distinction between cross-platform generalization and estimation of platform-level trends. A classifier can perform well on both test sets while making different systematic errors at each source. For example, a modest rate of false negatives for complaints may have a larger effect on a platform where complaints are rarer. If the output is used to compare the share of negative reviews, use a stratified human audit to estimate classification errors in each platform and period. Report the annotated sample size, the uncertainty around the adjusted proportion, and the sampling rules that define the population of posted reviews.

Documenting a dataset and documenting a model serve different readers. A dataset description should state how posts were selected, how annotation was performed, and which populations or platforms are represented. Datasheets provide a general structure for recording those decisions [17]. A model report should state which version was evaluated, the intended uses, performance across important slices, and limitations of external validity. The model-card framework offers a starting point for that report [18]. A research team can keep both records short by linking to the exact benchmark split and recording each processing decision once.

Three research questions follow directly from the evidence. The first is whether matched polarity annotations reduce the observed gap between food-delivery and product reviews. The second is whether modern encoders and prompted language models fail on the same reviews after annotation disagreements are accounted for. The third is whether a classifier's error profile remains stable when evaluated on reviews from a later period. Answering these questions with fixed protocols would establish which performance differences matter for practical review analysis and which reflect the benchmark design.

Future work can also compare active annotation with probability sampling. Active selection can focus labeling effort on difficult comments, as explored for Persian sentiment by Ashrafi Asli et al. [5]. A sample optimized for training, however, may overrepresent ambiguous text. Use a separate probability or stratified sample for estimating population performance or reporting the proportion of negative reviews. That separation allows an organization to improve a classifier while preserving an interpretable estimate of how it behaves in the review stream.

11. LIMITATIONS AND CONCLUSION

The review selects studies with relevant and inspectable benchmark evidence. It does not estimate the prevalence of particular evaluation practices across the field. Its conclusions concern Iranian Persian online product and service reviews and should be revisited for other Persian varieties, platforms, and annotation targets. The numerical examples illustrate interpretation within their original studies. No models were retrained, and the cited results were not independently replicated.

The reviewed studies show that supervised encoders and prompted large language models can classify sentiment in Persian online reviews. Which model performs better depends on the benchmark and protocol. A score is most informative when the reader knows what a label represents, how the test examples were isolated, and which types of error remain. For consumer research, a common label definition and a genuinely held-out target platform are the priorities for testing whether a classifier can be reused across sources.

Author Contributions

The author was solely responsible for the conception and design of the review, literature identification and evaluation, synthesis and interpretation of the reviewed evidence, development of the proposed reporting framework, and preparation of the manuscript. The author read and approved the final manuscript.

Use of Generative AI

During the preparation of this work, the author used generative AI tools to assist with improving the language, clarity, and readability of the manuscript. The author reviewed and edited the content as necessary and takes full responsibility for the accuracy, originality, and overall content of the publication.

Acknowledgments

The author would like to acknowledge the researchers whose published studies and datasets contributed to the evidence reviewed and synthesized in this article.

Competing Interests

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Considerations

This study was a focused narrative review based on previously published literature and publicly available research materials. It did not involve the collection of new data from human participants or animals and therefore did not require ethical approval or informed consent.

Data Availability

No new primary data were created or collected for this study. The evidence synthesized in this review was derived from previously published studies and publicly available research materials. Additional information regarding the review process may be available from the corresponding author upon reasonable request.

REFERENCES

[1] Hosseini, P., Ramaki, A. A., Maleki, H., Anvari, M., & Mirroshandel, S. A. (2018). SentiPers: A sentiment analysis corpus for Persian. arXiv:1801.07737. https://arxiv.org/abs/1801.07737

[2] PourMostafa Roshan Sharami, J., Abbasi Sarabestani, P., & Mirroshandel, S. A. (2020). DeepSentiPers: Novel deep learning models trained over proposed augmented Persian sentiment corpus. arXiv:2004.05328. https://arxiv.org/abs/2004.05328

[3] Hooshvare Lab. (n.d.). Persian sentiment analysis [Dataset documentation]. https://hooshvare.github.io/docs/datasets/sa

[4] Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based model for Persian language understanding. Neural Processing Letters, 53, 3831–3847. https://doi.org/10.1007/s11063-021-10528-4

[5] Ashrafi Asli, S. A., Sabeti, B., Majdabadi, Z., Golazizian, P., Fahmi, R., & Momenzadeh, O. (2020). Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian. Proceedings of LREC 2020, 2855–2861. https://aclanthology.org/2020.lrec-1.348/

[6] Masumi, M., Majd, S. S., Shamsfard, M., & Beigy, H. (2025). FaBERT: Pre-training BERT on Persian blogs. Proceedings of W-NUT 2025, 85–96. https://aclanthology.org/2025.wnut-1.10/

[7] Basiri, M. E., & Kabiri, A. (2020). HOMPer: A new hybrid system for opinion mining in the Persian language. Journal of Information Science, 46(1), 101–117. https://doi.org/10.1177/0165551519827886

[8] Shumaly, S., Yazdinejad, M., & Guo, Y. (2021). Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings. PeerJ Computer Science, 7, e422. https://doi.org/10.7717/peerj-cs.422

[9] SadraeiJavaheri, M., Moghaddaszadeh, A., Molazadeh, M., Naeiji, F., Aghababaloo, F., Rafiee, H., Amirmahani, Z., Abedini, T., Sheikhi, F. Z., & Salehoof, A. (2024). TookaBERT: A step forward for Persian NLU. arXiv:2407.16382. https://arxiv.org/abs/2407.16382

[10] Arnob, N. M. K., & Mahi, A. B. S. (2026). One language, three of its voices: Evaluating multilingual LLMs across Persian, Dari, and Tajiki on translation and understanding tasks. Proceedings of SilkRoadNLP 2026, 98–104. https://doi.org/10.18653/v1/2026.silkroadnlp-1.10

[11] Shangipour Ataei, T., Darvishi, K., Javdan, S., Minaei-Bidgoli, B., & Eetemadi, S. (2022). Pars-ABSA: A manually annotated aspect-based sentiment analysis benchmark on Farsi product reviews. Proceedings of LREC 2022, 7056–7060. https://aclanthology.org/2022.lrec-1.763/

[12] Amiri, F., Scerri, S., & Khodashahi, M. (2015). Lexicon-based sentiment analysis for Persian text. Proceedings of RANLP 2015, 9–16. https://aclanthology.org/R15-1002/

[13] Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of ACL 2007, 440–447. https://aclanthology.org/P07-1056/

[14] Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. Proceedings of ACL 2022, 8424–8445. https://aclanthology.org/2022.acl-long.577/

[15] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. Proceedings of ACL 2018, 1383–1392. https://aclanthology.org/P18-1128/

[16] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML 2017, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[17] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

[18] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of FAT* 2019, 220–229. https://doi.org/10.1145/3287560.3287596