Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 521-532.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Enhancing Minority Class Detection in Persian E-commerce Sentiment Analysis via Focal Loss and Explainable AI

Kiana Rezaei Jafari[1], Omid Mahdi Ebadati E.[1*] and Hamza Khastar[2]

1. Department of Operation Management and Information Technology, Faculty of Management, Kharazmi, Tehran, Iran.
2. Department of Business Management, Faculty of Management, Kharazmi, Tehran, Iran.

Article Info

Article History:
Received 23 April 2026
Revised 08 July 2026
Accepted 01 August 2026

DOI:10.22044/jadm.2026.17616.2911

Keywords:
Sentiment Analysis, Persian Natural Language Processing, Class Imbalance, Focal Loss, Explainable Artificial Intelligence, E-commerce Reviews

**Corresponding author: ebadati@khu.ac.ir (O.M Ebadati E.)*

Abstract

The rapid growth of e-commerce has led to an increasing volume of Persian user reviews containing valuable insights into products and services. Sentiment analysis enables the automatic extraction of sentiment polarity from such data; however, Persian sentiment analysis remains underexplored, particularly in real-world e-commerce contexts. In addition, class imbalance in sentiment datasets poses a major challenge, often leading to biased models that underperform on minority classes. In this study, we investigate Persian e-commerce sentiment analysis by comparing classical machine learning models, including Logistic Regression and SVM, with transformer-based models, namely ParsBERT and ParsRoBERTa. To mitigate the impact of class imbalance, we investigate Focal Loss against the standard Cross-Entropy Loss. Furthermore, we employ Integrated Gradients within an Explainable Artificial Intelligence (XAI) framework to improve model interpretability and analyze feature contributions. Experimental results on the Digikala dataset demonstrate that ParsBERT trained with Focal Loss achieves the best performance, reaching a Balanced Accuracy of 88.64% and an AUC-ROC of 95.91%. The findings highlight the effectiveness of combining imbalance-aware loss functions with transformer-based architectures for improving minority class detection in Persian sentiment analysis.

1. Introduction

In recent years, e-commerce has emerged as one of the fastest-growing industries worldwide, driven by the rapid expansion of online platforms, digital marketplaces, and social media [1]. As a result, an enormous volume of user-generated content is continuously produced, through which customers share their opinions and experiences regarding products and services [2]. These online reviews play a crucial role in shaping consumer purchasing decisions, as potential buyers increasingly rely on the experiences of others before making informed choices [3]. Previous studies have also shown that even subtle variations in sentiment expression can significantly influence customer behavior [4]. Therefore, analyzing such textual data is not only valuable for consumers but also essential for businesses seeking to improve product quality, enhance service delivery, and strengthen customer satisfaction [5]. In this context, sentiment analysis, as a fundamental task in natural language processing (NLP), provides an effective approach for automatically extracting sentiment polarity from large-scale textual data. Its applications span various domains, including marketing, customer relationship management, and consumer behavior analysis, particularly in e-commerce environments where understanding customer feedback is critical for decision-making. Despite these advancements, most existing research in sentiment analysis has primarily focused on high-resource languages such as English, while low-resource languages like Persian remain comparatively underexplored [6].

This limitation is more pronounced due to linguistic complexity, the limited availability of annotated resources, and domain-specific challenges associated with Persian text processing. Consequently, the development of robust sentiment analysis models tailored specifically to Persian remains an important and open research problem. In addition to language-related limitations, real-world e-commerce datasets are often highly imbalanced, with positive reviews significantly outnumbering negative ones [7]. Such imbalance can bias learning algorithms toward the majority class, reducing their effectiveness in detecting minority-class instances, particularly negative or critical opinions. Moreover, many existing studies rely heavily on overall accuracy, which is not sufficient for evaluating performance in imbalanced scenarios. Therefore, the use of more appropriate evaluation metrics is essential for obtaining a reliable and unbiased assessment of model performance [8]. Furthermore, although recent advances in deep learning and transformer-based models have significantly improved sentiment classification performance, these models are often considered black-box systems, as their internal decision-making processes are difficult to interpret [9]. This lack of transparency limits their applicability in real-world business environments, where interpretability and trust are essential. To address this issue, an Explainable Artificial Intelligence (XAI) method has been developed to provide insights into model behavior and highlight the most influential features contributing to model predictions [10].

Motivated by these challenges, this study aims to improve minority class detection in Persian sentiment analysis under imbalanced e-commerce data conditions. Specifically, it focuses on addressing data imbalance through an effective learning strategy while ensuring a reliable and comprehensive evaluation using appropriate metrics. In addition, an XAI technique is employed to enhance model interpretability and analyze the contribution of input features to model predictions. The main contributions of this study are summarized as follows:

- A unified comparative analysis of classical machine learning models and transformer-based models under consistent experimental settings.
- A systematic investigation of Focal Loss versus Cross-Entropy Loss, demonstrating its effectiveness in improving minority class detection.
- An Integrated Gradients-based analysis to investigate how different loss functions influence model predictions and feature attribution.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the proposed methodology. Section 4 discusses the experimental results and analysis. Finally, Section 5 concludes the study and outlines potential directions for future research.

2. Related Works

Sentiment analysis in the e-commerce domain has become one of the most important applications of NLP, attracting significant attention from both academia and industry in recent years. With the rapid growth of user-generated content on online platforms, the demand for accurate and automated sentiment classification systems has increased substantially. Existing research has explored this problem from multiple perspectives, ranging from traditional machine learning approaches to deep learning architectures, transformer-based models, and XAI techniques.

In traditional machine learning approaches, algorithms such as Logistic Regression, SVM, and Random Forests have been widely used for sentiment classification tasks. For example, Akter et al. [11] evaluated several models, including Logistic Regression, Random Forest, SVM, LSTM, and BERT, on customer review datasets. Their results showed that BERT achieved the highest performance with an accuracy of 94.20% and an AUC-ROC of 97%, confirming the effectiveness of transformer-based models in capturing contextual semantic relationships. Similarly, Singh et al. [12] proposed a sentiment classification framework for Kindle Store book reviews and addressed class imbalance using undersampling techniques. Although their approach improved the performance of classical models, it also led to the loss of potentially informative samples. With the advancement of deep learning, particularly transformer-based architectures, sentiment analysis performance has improved significantly. Taneja et al. [13] applied DistilBERT to classify reviews in the online fashion domain, achieving an accuracy of 96% and an F1-score of 79%. Their study highlighted that, while traditional models such as Bi-LSTM were trained using both oversampling and undersampling techniques, transformer-based models were not explicitly optimized for class imbalance. In another study, Harunasir et al. [14] compared several models, including Multinomial Naive Bayes, Random Forest, LSTM, and CNN, for identifying dissatisfied reviews in Amazon datasets, reporting that LSTM achieved the best

performance with 97% accuracy.

Sentiment analysis in low-resource languages such as Persian and Dari presents additional challenges due to limited linguistic resources and higher structural complexity. Muradi et al. [15] investigated emotion detection in Dari using the ArmanEmo dataset and compared mBERT and ParsBERT models. Their results showed that ParsBERT outperformed mBERT, achieving 86.4% accuracy compared to 81.2%, and data augmentation was used to partially address class imbalance. Similarly, Ariai et al. [16] proposed an aspect-based sentiment analysis model for Digikala reviews based on ParsBERT with lexical enhancements, achieving an accuracy of 88.2% and an F1-score of 61.7%. In a large-scale study, Shumaly et al. [17] collected over three million Digikala reviews and applied FastText embeddings combined with CNN architectures, achieving an AUC-ROC of 99.6% and an F1-score of 95.6%, demonstrating the effectiveness of embedding-based deep learning methods in sentiment classification tasks.

Beyond studies specifically focusing on Persian sentiment analysis, recent advances in Persian NLP have been marked by the emergence of increasingly sophisticated monolingual transformer variants and specialized pre-training strategies tailored to the linguistic characteristics of Persian. These efforts aim to better capture the linguistic diversity of Persian, including informal language, domain-specific vocabulary, and its rich morphological characteristics [18]. For instance, FaBERT was introduced to explicitly capture the nuances of informal and colloquial Persian by pre-training on large-scale web and blog corpora, demonstrating improved performance in social media sentiment analysis and irony detection [19]. More recently, Tiny-ParsBERT was proposed as a lightweight transformer architecture specifically designed for efficient Persian sentiment analysis. By reducing the number of transformer layers and incorporating adversarial training, the model substantially decreases computational complexity while maintaining high predictive performance [20]. Furthermore, recent advancements have introduced scaled encoder models such as Tooka-SBERT, which utilizes modern pre-training techniques and specialized tokenizers to achieve improved performance over previous Persian transformer models across multiple Persian natural language understanding benchmarks [21]. Concurrently, the field has expanded into generative architectures through models like PersianMind, which adapts foundational large language model frameworks to Persian-specific instruction tuning and complex bilingual reasoning tasks by expanding token vocabularies [22]. Such generative trends are further exemplified by recent native frameworks such as FarsEval-PKBETS, which have been evaluated on specialized Persian benchmarks to improve cultural and linguistic alignment [23]. Collectively, these recent developments highlight the rapid evolution of Persian-specific transformer architectures and pre-training strategies, providing a stronger foundation for downstream NLP applications, including sentiment analysis, text classification, and domain-specific language understanding.

In addition to performance improvements, the interpretability of sentiment analysis models has become an increasingly important research direction. Muradi et al. [15] applied both LIME and SHAP to interpret ParsBERT predictions in Persian sentiment classification and identified key influential words affecting model decisions. Jahan et al. [24] extended BERT-based models for multilingual sentiment analysis and used LIME to improve interpretability, achieving an overall accuracy of 92%. Similarly, Jahin et al. [25] applied LIME and SHAP to explain transformer-based sentiment classification, while Krawczyk et al. [26] reported that Bi-LSTM achieved strong performance with an AUC-ROC of 98.20% and an accuracy of 93.53%, highlighting the importance of incorporating XAI techniques into deep learning-based sentiment analysis systems.

Despite these advancements, several research gaps remain. First, although Persian-specific transformer architectures have demonstrated strong performance, their behavior under class-imbalanced conditions has not been extensively studied. Most existing works rely on data-level techniques such as undersampling, oversampling, or data augmentation, while limited attention has been given to loss-function-based approaches tailored for imbalanced learning. Second, Focal Loss, which emphasizes the learning process on hard-to-classify examples while down-weighting easy samples, has rarely been explored in Persian sentiment analysis tasks, despite its potential to improve minority class detection. Third, although XAI techniques such as LIME and SHAP are commonly used, more advanced methods such as Integrated Gradients remain underexplored in the context of Persian NLP.

To address these limitations, this study investigates the impact of Focal Loss on transformer-based sentiment analysis models for Persian e-commerce reviews. Additionally, Integrated Gradients is employed to provide deeper interpretability of model predictions and to analyze the effect of

different loss functions on model behavior. By combining imbalance-aware learning strategies with an XAI technique, this work aims to improve both the performance and interpretability of sentiment analysis systems in low-resource language settings.

3. Methodology

In this study, we propose a systematic framework for sentiment analysis of user reviews in the e-commerce domain. The overall architecture of the proposed framework is illustrated in Figure 1 and consists of five main stages: data collection, preprocessing, modeling, training, and evaluation. XAI is integrated into the framework to analyze model behavior and enhance the interpretability of predictions.

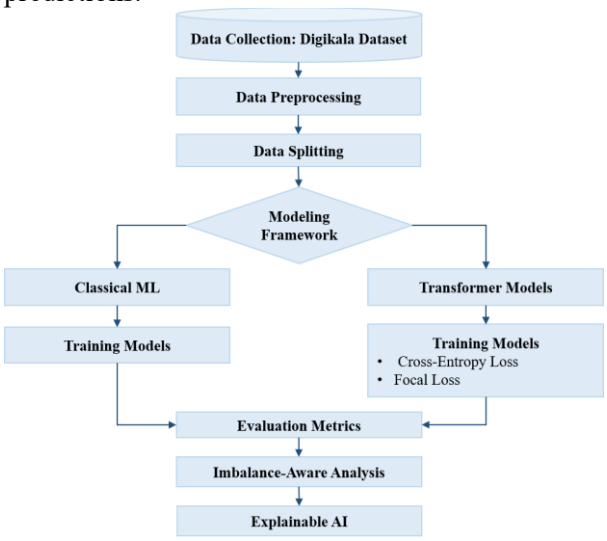

Figure 1. Overview of the proposed framework for sentiment analysis of Persian e-commerce reviews.

Data Collection

The dataset used in this study is SentiPers, developed by Hooshvare Lab and publicly available for research purposes [27]. SentiPers is a benchmark dataset for Persian sentiment analysis containing approximately 15,500 user reviews extracted from the Digikala platform. The dataset covers multiple product categories, including electronics, mobile phones, and accessories. Reviews are annotated using a five-point Likert scale, where labels 2 and 1 represent very positive and positive sentiments; 0 indicates neutral sentiment, and -1 and -2 correspond to negative and very negative sentiments, respectively.

3.1. Data Preprocessing

Since the dataset is already cleaned and manually annotated, no aggressive preprocessing steps such as stemming, lemmatization, or stopword removal were applied. This decision was made to preserve the original semantic and syntactic structure of user-generated reviews, ensuring that models learn from realistic textual patterns.

For this study, the sentiment classification task was reformulated as a binary problem. Specifically, neutral samples (label 0) were removed, while positive labels (1 and 2) were merged into a single positive class and negative labels (-1 and -2) into a single negative class.

To ensure robust evaluation, the dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling. This approach preserves the original class distribution across all subsets and reduces potential sampling bias. The final dataset exhibits a noticeable class imbalance, with the positive class containing 7,959 samples and the negative class containing 1,786 samples, as illustrated in Figure 2. This imbalance motivates the use of appropriate evaluation metrics and loss considerations.

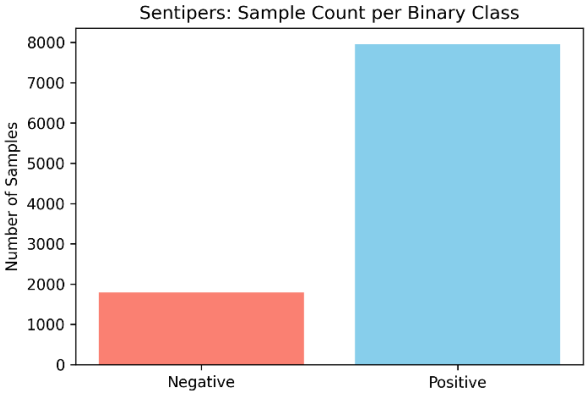

Figure 2. Distribution of samples across positive and negative classes in the binary sentiment classification task, highlighting the inherent class imbalance.

Models and Experimental Setup

The proposed framework evaluates two categories of models: classical machine learning models and transformer-based architectures.

3.1.1. Classical Models

Logistic Regression and SVM are employed as baseline models due to their established performance in text classification tasks and computational efficiency. Text representations are constructed using TF-IDF with unigrams and bigrams to capture both local and contextual lexical features. To mitigate the effects of class imbalance, balanced class weighting is incorporated into both models, assigning greater importance to misclassification errors associated with the minority class. Hyperparameter optimization is performed using GridSearchCV on the validation set. The optimal regularization strengths are determined based on the Macro F1-score, with the LR and SVM models achieving their best performance at regularization values of 10 and 0.5, respectively.

3.1.2. Transformer-Based Models

Transformer-based architectures have become the dominant paradigm in NLP due to their ability to capture contextual dependencies using the self-attention mechanism. Unlike sequential models such as RNNs, transformers process tokens in parallel and dynamically assign importance weights to words in a sequence [28]. The attention mechanism is defined as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K and V represent query, key, and value matrices, and d_k denotes the dimensionality of the key vectors. This mechanism enables the model to learn contextual relationships by assigning higher weights to more relevant tokens.

Among transformer-based models, BERT (Bidirectional Encoder Representations from Transformers) introduces a bidirectional training strategy that considers both left and right context simultaneously [29] RoBERTa further improves upon BERT by optimizing pretraining procedures, removing the Next Sentence Prediction objective, and using larger training corpora and batch sizes, leading to improved performance across NLP tasks [30] For Persian language modeling, ParsBERT and ParsRoBERTa are employed. These models are adaptations of BERT and RoBERTa, respectively, and are pre-trained on large-scale Persian corpora to better capture linguistic characteristics specific to the Persian language [31]. Both models are fine-tuned on the target dataset under identical training conditions to ensure a fair comparison. Training is performed using the AdamW optimizer with weight decay for improved generalization. A learning rate warm-up strategy is applied at the beginning of training, followed by gradual decay. Early stopping is used with a patience of 2 epochs based on validation Macro F1-score to prevent overfitting. The best-performing model weights are restored during training. The hyperparameter settings employed throughout all transformer-based experiments are summarized in Table 1.

Table 1. Hyperparameter settings for transformer-based models used in this study.

Parameter	ParsBERT	ParsRoBERTa
Learning Rate	1e-5	3e-5
Batch Size	16	16
Epoch	5	10
Patience	2	2
Max Length	128	128

Loss Functions

To address class imbalance, this study investigates Focal Loss as the primary learning objective. Focal Loss is designed to reduce the contribution of well-classified samples while focusing training on hard-to-classify instances, thereby improving learning on minority classes [32]. For comparison, Cross-Entropy Loss is used as a baseline. Cross-Entropy Loss treats all samples equally and measures the divergence between predicted probabilities and true labels [33], defined as:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \quad (2)$$

where y_i is the true label and p_i is the predicted probability.

Focal Loss introduces a modulation factor to down-weight easy examples and focus on hard ones, defined as:

$$L_{FL} = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (3)$$

where p_t is the predicted probability of the true class, α is a balancing factor, and γ controls the focusing strength. In this study, γ is set to 2 and α is determined based on the class distribution.

3.5. Evaluation Metrics

To evaluate model performance under class imbalance conditions, Balanced Accuracy, Macro F1-score, and AUC-ROC are used as primary metrics. In addition, Accuracy, Precision, Recall and F1-score are reported for completeness.

Balanced accuracy is defined as the average recall across all classes, making it suitable for imbalanced datasets:

$$Balanced\ Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Re}call_i \quad (4)$$

Macro F1-score computes the F1-score independently for each class and averages them equally.

$$Macro\ F1 = \frac{1}{N}\sum_{i=1}^{N} F1_i \quad (5)$$

3.6. Explainable Artificial Intelligence Using Integrated Gradients

To interpret model predictions and analyze the impact of Focal Loss on classification behavior, the Integrated Gradients (IG) method is used [34]. IG attributes predictions to input features by integrating gradients along a path from a baseline input to the actual input. Formally, Integrated Gradients is defined as:

$$IG(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} d\alpha \qquad (6)$$

where x is the input, x' is the baseline, $f(x)$ is the model output, and α is the interpolation factor. The method is implemented using the Captum library, and 50 interpolation steps are used for numerical approximation. Results are visualized as token-level heatmaps, where:

- **Green indicates positive contribution,**
- **Red indicates negative contribution,**
- **White indicates negligible contribution.**

4. Experiments and Results

The experimental results indicate that all evaluated models achieve strong overall performance, with Accuracy consistently exceeding 89% across all experimental settings. However, given the pronounced class imbalance in the dataset, Accuracy alone is not a reliable evaluation metric. Therefore, more robust measures, including Balanced Accuracy, Macro F1-score, and AUC-ROC, are reported in Table 2 to provide a more comprehensive assessment. As shown in Table 2, classical machine learning models provide competitive baseline performance. In particular, SVM consistently outperforms Logistic Regression. SVM achieves a Balanced Accuracy of 82.64% and an AUC-ROC of 93.45%, indicating its stronger generalization capability compared with Logistic Regression.

Table 2. Comparison of proposed model performance across evaluation metrics.

Models	Class	Precision	Recall	F1-score	Accuracy	Balanced Accuracy	Macro F1-score	AUC-ROC
LR	P	92.55	94.64	93.58	89.40	80.34	81.56	92.64
	N	73.44	66.04	69.55				
SVM	P	93.53	94.47	94	90.15	82.64	83.26	93.45
	N	74.22	70.90	72.52				
ParsRoBERTa + CE	P	93	96	95	91	82.88	84.51	94.21
	N	80	70	74				
ParsRoBERTa + FL	P	94	95	95	91	83.95	84.70	94.39
	N	77	73	75				
ParsBERT + CE	P	95	95	95	92	86.40	86.24	95.73
	N	77	78	78				
ParsBERT + FL	P	97	91	94	90	**88.64**	85.32	**95.91**
	N	69	**86**	77				

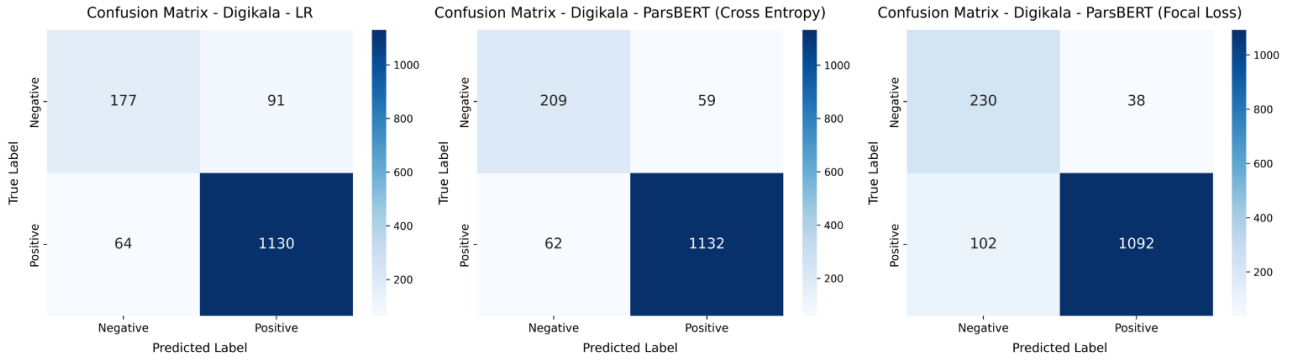

Figure 3. Confusion Matrices for (left to right) Logistic Regression, ParsBERT trained with Cross-Entropy Loss, and ParsBERT trained with Focal Loss, illustrating differences in classification performance and minority class detection.

In contrast, transformer-based models significantly outperform classical approaches. Among them, ParsBERT demonstrates the best overall performance under both loss settings. When trained with Cross-Entropy Loss, ParsBERT achieves a Balanced Accuracy of 86.40% and an AUC-ROC of 95.73%. Replacing Cross-Entropy Loss with Focal Loss further improves performance, increasing Balanced Accuracy to 88.64% and AUC-ROC to 95.91%. These results confirm the effectiveness of Focal Loss in improving sensitivity toward minority-class samples under imbalanced conditions. To further investigate the effect of the focusing parameter in Focal Loss, a sensitivity analysis was conducted using the best-performing model. Since γ controls the degree to which the loss function emphasizes hard-to-classify samples, its value can substantially influence classification performance on imbalanced datasets. To isolate the effect of this hyperparameter, all experimental settings, including the dataset, data split, optimizer, learning rate, batch size, and training configuration, were kept identical, while only the value of the focusing

parameter was varied from 0 to 3. The results are summarized in Table 3.

Table 3. Sensitivity analysis of the focusing parameter (γ) for ParsBERT with Focal Loss.

γ	Accuracy	Balanced Accuracy	Macro F1-score	AUC-ROC	Recall (Negative)
0	88.71	86.87	83.01	94.44	83.96
1	88.71	86.58	82.93	94.58	83.21
2	**90**	**88.64**	**85.32**	**95.91**	**86**
3	88.37	86.66	82.60	93.97	83.96

As shown in Table 3, the model exhibits moderate sensitivity to the focusing parameter. Although the overall performance remains relatively stable across the examined values, the setting of 2 consistently achieves the best results, yielding an Accuracy of 90%, a Balanced Accuracy of 88.64%, a Macro F1-score of 85.32%, and an AUC-ROC of 95.91%. In addition, this configuration achieves the highest Recall for the minority class (86%), indicating its superior ability to identify difficult samples. Compared with the optimal setting, smaller values place less emphasis on difficult samples, whereas a larger value slightly degrades overall performance across most evaluation metrics. Overall, these findings indicate that a focusing parameter value of 2 provides the most suitable configuration for the dataset and experimental settings considered in this study, offering the best trade-off between minority-class detection and overall classification performance.

Additional insights into the effect of Focal Loss can be obtained from the confusion matrix analysis of Logistic Regression and ParsBERT under different loss functions, as illustrated in Figure 3. Logistic Regression shows a relatively high number of misclassified negative samples, indicating limited sensitivity toward the minority class. In contrast, ParsBERT trained with Focal Loss reduces the number of negative reviews incorrectly classified as positive, indicating improved minority-class detection, demonstrating enhanced capability in detecting negative reviews.

The observed improvement can be attributed to Focal Loss, which reduces the contribution of easily classified samples while emphasizing difficult and misclassified instances. This mechanism is particularly beneficial for the imbalanced dataset, where negative reviews represent the minority class. Combined with ParsBERT's contextual representations, Focal Loss encourages the model to learn more discriminative features for minority-class reviews. The improvement in minority-class Recall and Balanced Accuracy indicates a more balanced classification behavior. Moreover, the high AUC-ROC suggests that improved minority-class sensitivity does not substantially compromise overall class discrimination.

Despite these improvements, a clear trade-off is observed. While Focal Loss substantially enhances Recall for the minority class, it slightly increases misclassified positive reviews in the majority class. Specifically, the Recall of the negative class increases from 78% to 86% in ParsBERT, indicating improved detection of hard-to-classify samples. This behavior reflects a more conservative decision boundary, where the model prioritizes minority-class detection at the expense of a slight reduction in precision. A detailed inspection of confusion matrices further confirms that Focal Loss reduces misclassified negative reviews, while slightly increasing misclassified positive reviews. This explains the improvement in Balanced Accuracy, despite minor fluctuations in Macro F1-score. From a practical perspective, this trade-off is desirable in e-commerce sentiment analysis. In real-world applications, failing to detect negative reviews may have a more severe impact on decision-making and customer satisfaction than incorrectly flagging positive reviews. Therefore, models optimized with Focal Loss are more suitable for deployment in imbalance-sensitive environments. The overall comparison of model performance across evaluation metrics is further illustrated in Figure 4.

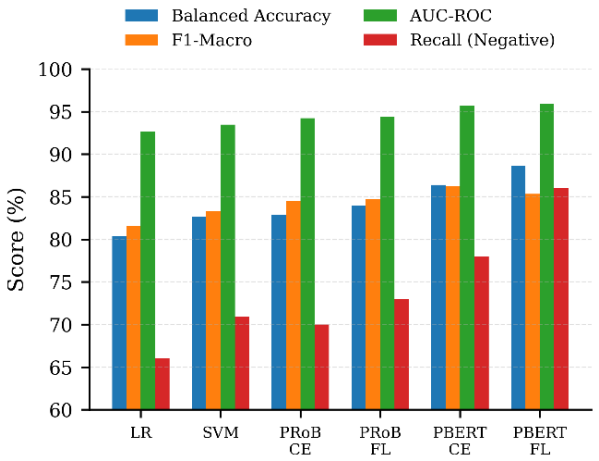

Figure 4. Overall comparison of model performance across key evaluation metrics. PRoB and PBERT represent ParsRoBERTa and ParsBERT models, respectively, while CE and FL refer to Cross-Entropy Loss and Focal Loss, respectively.

Finally, the AUC-ROC curves, presented in Figure 5, provide an aggregate measure of class separability across different classification thresholds. A higher AUC-ROC indicates better discrimination between positive and negative classes. In this study, ParsBERT achieves the highest AUC-ROC score, confirming its superior

discriminative capability compared with both classical and other transformer-based models.

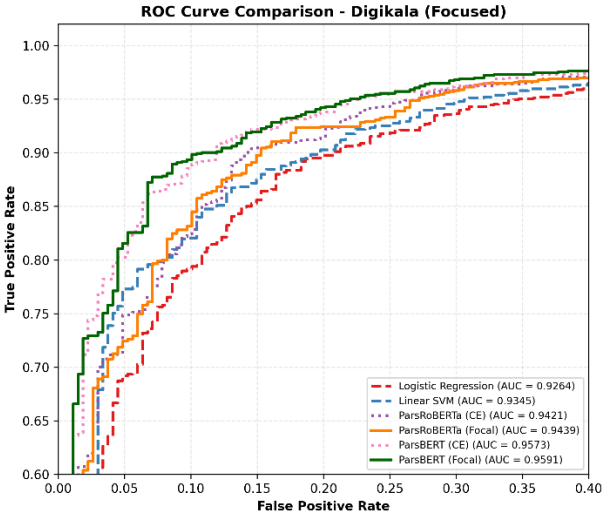

Figure 5. ROC curve comparison of all models with corresponding AUC scores

Overall, ParsBERT outperforms the best classical baseline (SVM) by approximately 5.96 percentage points in Balanced Accuracy and 2.46 percentage points in AUC-ROC. This demonstrates the superiority of contextual representation learning over traditional TF-IDF-based feature engineering in sentiment analysis tasks. Although transformer-based models consistently achieve superior classification performance, they generally require substantially higher computational resources than classical machine learning approaches. To quantify this trade-off, we measured the training time, total inference time, and average inference latency per sample for all evaluated models under the same experimental environment. The results are summarized in Table 4.

Table 4. Computational cost comparison of the evaluated models.

Model	Training Time (s)	Total Inference Time (s)	Inference Latency per sample (ms)
LR	1.34	0.124	0.085
SVM	0.59	0.068	0.046
ParsRoBERTa + CE	714.48	9.386	6.420
ParsRoBERTa + FL	709.31	9.520	6.512
ParsBERT + CE	420.10	10.152	6.944
ParsBERT + FL	421.31	10.269	7.024

As expected, the classical models (LR and SVM) exhibit considerably lower computational cost, requiring less than 2 seconds for training and less than 0.13 seconds for inference over the entire test set. In contrast, transformer-based models require substantially longer training and inference times due to their larger model size and contextual representation learning. However, the results indicate that incorporating Focal Loss introduces only a negligible computational overhead compared with the standard Cross-Entropy Loss. For example, the training time of ParsBERT increases only marginally from 420.10 s to 421.31 s, while the average inference latency changes from 6.944 ms to 7.024 ms per sample. Similar observations can be made for ParsRoBERTa, where the computational cost remains nearly unchanged under both loss functions. These findings demonstrate that the performance gains obtained by Focal Loss introduce negligible additional computational overhead relative to Cross-Entropy Loss, making it a practical choice for real-world Persian e-commerce sentiment analysis.

Explainable Artificial Intelligence Analysis Using Integrated Gradients

To further interpret the behavior of the ParsBERT model and analyze the impact of Focal Loss on decision-making, an XAI analysis was conducted using Integrated Gradients. This analysis focuses on understanding how the model's attribution pattern changes across different tokens under Cross-Entropy Loss and Focal Loss settings. The results of token-level attribution, along with corresponding predictions, are summarized in Table 5. The selected examples include correctly classified negative samples as well as misclassified cases introduced by Focal Loss, enabling a comparative analysis of model behavior under different loss functions. As shown in Table 5, clear differences can be observed between Cross-Entropy Loss and Focal Loss in terms of token attribution and prediction behavior.

In the first review, which contains both positive and sudden negative sentiment shifts, the model trained with Cross-Entropy Loss mainly focuses on early positive tokens such as "excellent," and "really," resulting in an incorrect positive prediction. In contrast, Focal Loss produces higher attribution for sentiment-reversing expressions such as "suddenly," "turned off," and "did not turn on," leading to correct classification of the negative sentiment.

Similarly, in the second review, which contains both positive descriptions and explicit product issues, the model trained with Cross-Entropy Loss is biased toward positive lexical cues such as "good features." However, Focal Loss assigns greater importance to failure-related and contrastive expressions such as "but," "overheating," "drawback," and "do not buy," resulting in correct negative classification. The heatmap-based visualization of these attribution patterns is shown

in Figure 6, providing a direct comparison between Cross-Entropy Loss and Focal Loss models. Despite improved sensitivity to negative sentiment, Focal Loss introduces certain error patterns. In one case, a review explicitly indicating “no problems” is misclassified as negative under Focal Loss due to oversensitivity to negation-related structures. In another case, strong positive sentiment is incorrectly classified as negative due to overemphasis on implicit contrast cues. These findings indicate that Focal Loss modifies the internal decision dynamics of ParsBERT by shifting feature attribution from superficial positive cues toward contextual and transition-based tokens. While this improves detection of minority-class samples, it may reduce robustness in linguistically simple or strongly positive cases.

Overall, the results confirm that Focal Loss is an effective strategy for handling class imbalance in sentiment analysis. At the same time, the findings highlight the importance of considering linguistic phenomena such as negation and contrast when designing loss-aware deep learning systems.

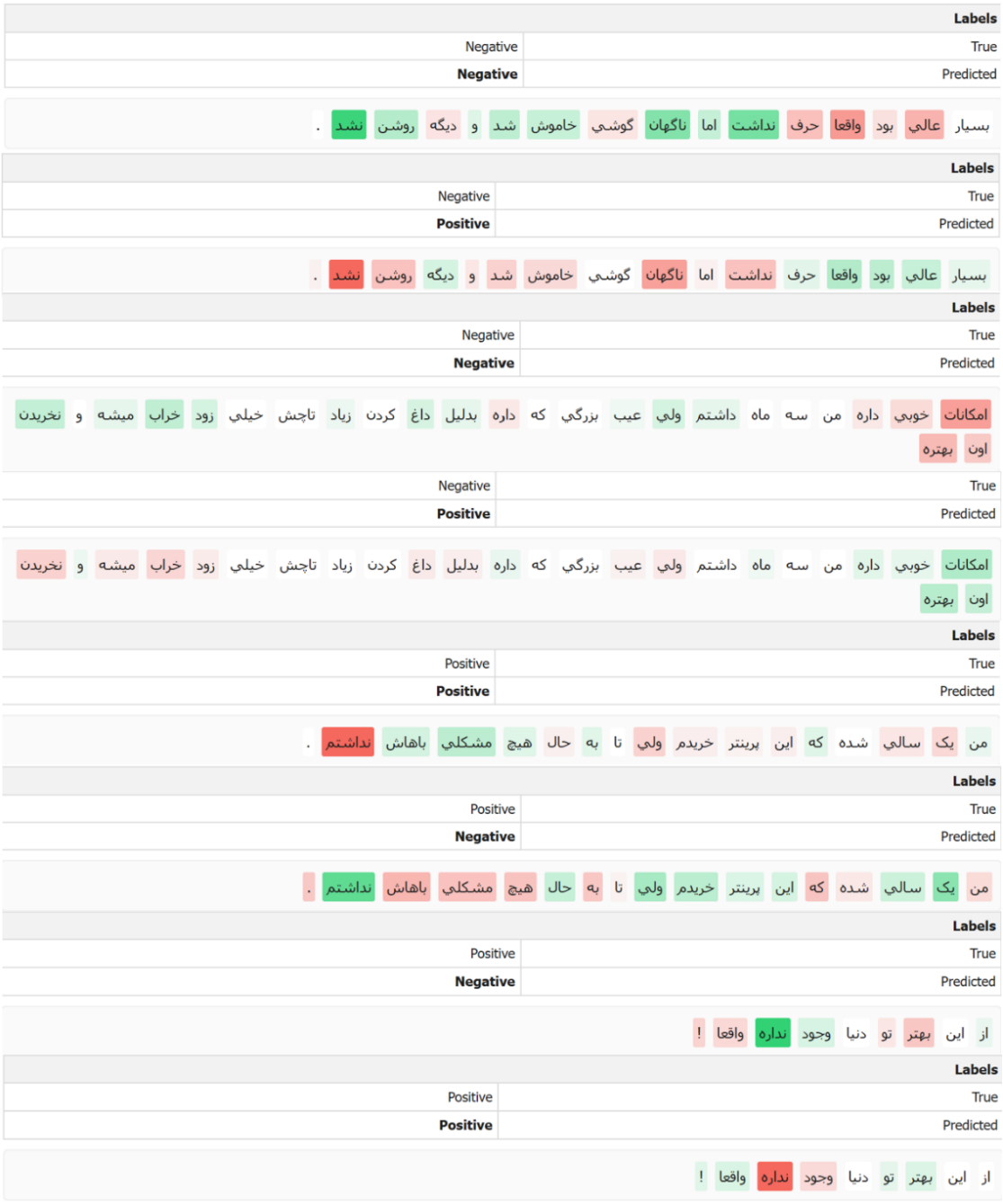

Figure 6. Comparison of Integrated Gradients heatmaps for Cross-Entropy Loss (CE) and Focal Loss (FL) across four representative test reviews. Each review is shown under both loss functions with its corresponding prediction and token-level attributions. The first two negative reviews are correctly classified by FL but misclassified by CE, whereas the last two positive reviews are correctly classified by CE but misclassified by FL.

Table 5. Integrated Gradients-based explanation of selected review cases under Cross-Entropy Loss and Focal Loss with ParsBERT model.

Review	CE (Focused tokens)	CE Prediction	FL (Focused tokens)	FL Prediction	True label
Persian: بسیار عالی بود واقعا حرف نداشت اما ناگهان گوشی خاموش شد و دیگه روشن نشد. English: It was really excellent and flawless, but suddenly the phone turned off and did not turn on again.	Persian: بسیار، عالی، واقعا English: excellent, really	Positive	Persian: ناگهان، خاموش شد، روشن نشد English: Suddenly, turned off, did not turn on	Negative	Negative
Persian: امکانات خوبی داره من سه ماه داشتم ولی عیب بزرگی که داره به دلیلی داغ کردن زیاد تاچش زود خراب میشه و نخرید اون بهتره. English: It has good features; I used it for three months, but its major drawback is overheating and touchscreen failure, and it is better not to buy it.	Persian: امکانات، خوبی، بهتره English: Good, features, better	Positive	Persian: ولی، عیب، داغ شدن، نخریدن English: But, overheating, drawback, do not buy	Negative	Negative
Persian: من یکسالی شده که این پرینتر خریدم ولی تا به حال هیچ مشکلی باهاش نداشتم. English: I have had this printer for a year, but so far I have not had any problems with it.	Persian: هیچ مشکلی نداشتم English: Have not had any problems	Positive	Persian: ولی، نداشتم English: But, have not had	Negative	Positive
Persian: از این تو دنیا بهتر وجود نداره واقعا! English: Honestly, there is nothing better than this in the world!	Persian: بهتر، واقعا! English: Nothing better, honestly	Positive	Persian: ندارد English: There is nothing	Negative	Positive

5. Conclusion and Future Work

This study investigated sentiment analysis in an imbalanced Persian e-commerce dataset, with a particular focus on addressing the limitations of conventional evaluation metrics such as Accuracy. Given the strong bias of Accuracy toward majority-class performance, more robust evaluation measures, including Balanced Accuracy, Macro F1-score, and AUC-ROC, were adopted to provide a more reliable assessment of model performance across both majority and minority classes. The experimental results demonstrate that transformer-based models consistently outperform classical machine learning approaches such as Logistic Regression and SVM. This improvement can be attributed to their ability to capture contextual semantics and long-range dependencies, which are essential for modeling complex linguistic patterns in real-world sentiment data. A key contribution of this study is the systematic investigation of Focal Loss within transformer-based architectures. In particular, the ParsBERT model trained with Focal Loss was evaluated against baseline models under a unified experimental setting. The results indicate that incorporating Focal Loss significantly enhances minority-class detection, leading to improved Balanced Accuracy and AUC-ROC, as well as a notable increase in Recall for negative reviews. Compared with the best-performing classical baseline (SVM), the proposed model achieves substantial improvements in both Balanced Accuracy and AUC-ROC. These findings highlight that, in imbalanced learning scenarios, loss function design can be as influential as architectural improvements in determining final model performance. Moreover, the computational cost analysis demonstrates that these performance improvements are achieved with only negligible additional overhead compared with the standard Cross-Entropy Loss. The training time and inference latency of ParsBERT remain virtually unchanged, indicating that the proposed approach offers an effective balance between classification performance and computational efficiency for practical Persian e-commerce sentiment analysis. To further interpret model behavior, an XAI analysis based on Integrated Gradients was conducted. The results reveal that ParsBERT trained with Cross-Entropy Loss tends to rely more on surface-level lexical cues, whereas the model trained with Focal Loss shifts its focus toward more context-dependent and semantically informative tokens. This attribution shift provides a possible explanation for the improved detection of minority-

class instances, particularly in cases involving mixed or conflicting sentiments. Overall, the findings suggest that Focal Loss not only improves classification performance but also meaningfully influences the learning dynamics of transformer-based models. This leads to more balanced decision boundaries and improved generalization in imbalanced and noisy real-world datasets. Despite these contributions, the study has certain limitations. The experiments are conducted on a single benchmark dataset (SentiPers extracted from Digikala), which, although widely used, does not fully capture the diversity of real-world Persian e-commerce data. As a result, the generalizability of the findings to other domains and datasets may be limited. Future research can extend this work in several important directions. First, the construction of large-scale and domain-diverse Persian sentiment datasets would significantly improve the robustness and external validity of evaluation results. Second, extending the current framework to fine-grained sentiment analysis tasks, such as Aspect-Based Sentiment Analysis (ABSA), could provide deeper insights at the product-feature level. Finally, future work could establish a unified benchmarking framework to systematically compare recent Persian-specific transformer models (e.g., Tooka-SBERT and Tiny-ParsBERT) together with emerging Persian Large Language Models (LLMs), such as PersianMind and FarsEval-PKBETS, under identical experimental settings for imbalanced Persian e-commerce sentiment analysis. Such investigations would provide a more comprehensive understanding of the trade-offs between predictive performance, computational efficiency, and model interpretability across next-generation Persian language models for Persian e-commerce sentiment analysis.

References

[1] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on sentiment analysis: Evolution of research methods and topics," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8469-8510, 2023.

[2] B. M. D. Abighail, Fachrifansyah, M. Firmanda, M. Anggreainy, Harvianto, and Gintoro, "Sentiment analysis e-commerce review," *Procedia Computer Science*, vol. 227, pp. 1039-1045, 2023.

[3] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, 2024.

[4] A. Daza, N. D. González Rueda, M. S. Aguilar Sánchez, W. F. Robles Espíritu, and M. E. Chauca Quiñones, "Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis, systematic literature review, challenges and future works," *International Journal of Information Management Data Insights*, vol. 4, no. 2, p. 100267, 2024.

[5] H. Huang, A. Asemi, and M. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90367-90382, 2023.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, 2022.

[7] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221-232, 2016.

[8] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 97-101.

[10] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv*:1702.08608, pp. 1-13, 2017.

[11] P. Akter, S. Hossain, M. Siddique, M. Ayub, A. Nath, P. Nath, M. Rasel, and M. Hassan, "Sentiment analysis of consumer feedback and its impact on business strategies by machine learning," *The American Journal of Applied Sciences*, vol. 7, no. 1, pp. 6-16, 2025.

[12] A. K. Singh, D. Sharma, M. Dhawan, S. Kumar, A. Chattaraj, and C. Gour, "Amazon kindle review sentiment analysis," in *2024 IEEE Recent Advances in Intelligent Computational Systems*, Thiruvananthapuram, India, 2024, pp. 1-6.

[13] K. Taneja, J. Vashishtha, and S. Ratnoo, "Transformer based unsupervised learning approach for imbalanced text sentiment analysis of e-commerce reviews," *Procedia Computer Science*, vol. 235, pp. 2318-2331, 2024.

[14] M. F. bin Harunasir, N. Palanichamy, S.-C. Haw, and K.-W. Ng, "Sentiment analysis of Amazon product reviews by supervised machine learning models," *Journal of Advances in Information Technology*, vol. 14, no. 4, pp. 857-862, 2023.

[15] M. Muradi, B. Hussain, E. R. Rhythm, and A. A. Rasel, "A comparative study of ParsBERT and mBERT in emotion recognition for Dari-Farsi text with Explainable AI," in *Proceedings of the 3rd International*

Conference on Computing Advancements, Dhaka, Bangladesh, 2022, pp. 399-405.

[16] F. Ariai, M. Tayefeh Mahmoudi, and A. Moeini, "Enhancing Aspect-based Sentiment Analysis with ParsBERT in Persian Language," *Journal of AI and Data Mining*, vol. 12, no. 1, pp. 1-14, 2024.

[17] S. Shumaly, M. Yazdinejad, and Y. Guo, "Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with FastText embeddings," *PeerJ Computer Science*, vol. 7, p. e422, 2021.

[18] Z. Bokaei, W. Magdy, and B. Webber, "Farsi natural language processing: A survey," *Information Processing & Management*, vol. 63, no. 8, p. 104912, 2026.

[19] M. Masumi, S. S. Majd, M. Shamsfard, and H. Beigy, "FABERT: Pre-training BERT on Persian blogs," in *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, Albuquerque, New Mexico, USA, 2024, pp. 85-96.

[20] M. Nooraee, H. Ghaffari, and F. Zarisfi Kermani, "Tiny-ParsBERT: An optimized hybrid model for efficient sentiment analysis in Persian texts," *The Journal of Supercomputing*, vol. 81, no. 7, p. 862, 2025.

[21] G. Zamaninejad, M. SadraeiJavaheri, F. Aghababaloo, H. Rafiee, M. M. Oskuee, and A. Salehoof, "Tooka-SBERT: Lightweight sentence embedding models for Persian," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Mumbai, India, 2024, pp. 2415-2425.

[22] P. Rostami, A. Salemi, and M. J. Dousti, "PersianMind: A cross-lingual Persian-English large language model," *arXiv preprint arXiv*:2401.06466, pp. 1-13, 2024.

[23] M. Shamsfard, Z. Saaberi, S. M. H. Hashemi, Z. Vatankhah, M. Ramezani, N. Pourazin, T. Zare, M. Azimi, S. Chitsaz, and S. Khoraminejad, "FarsEval-pkbets: A new diverse benchmark for evaluating Persian large language models," *arXiv preprint arXiv*:2502.10234, pp. 1-15, 2025.

[24] N. Jahan, J. Ahamed, and D. Nandi, "Enhancing e-commerce sentiment analysis with advanced BERT techniques," *International Journal of Information Engineering and Electronic Business*, vol. 17, no. 3, pp. 49-61, 2025.

[25] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," *Scientific Reports*, vol. 14, no. 1, p. 24882, 2024.

[26] K. Kyritsis, C. M. Liapis, I. Perikos, and M. Paraskevas, "Explainable sentiment analysis utilizing deep learning methods and LIME," in *Big Data and Data Science Engineering*, vol. 7, R. Lee, Ed. Cham, Switzerland :Springer Nature Switzerland, 2025, pp. 65-78.

[27] P. Hosseini, A. A. Ramaki, H. Maleki, M. Anvari, and S. A. Mirroshandel, "SentiPers: A sentiment analysis corpus for Persian," *arXiv preprint arXiv*:1801.07737, pp. 1-11, 2018.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000-6010, 2017.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics*: *Human Language Technologies*, vol. 1, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907*.11692, pp. 1-13, 2019.

[31] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "ParsBERT: Transformer-based model for Persian language understanding," *Neural Processing Letters*, vol. 53, no. 6, pp. 3831-3847, 2021.

[32] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.

[33] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, 1948.

[34] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, 2017, pp. 3319-3328.

بهبود تشخیص کلاس اقلیت در تحلیل احساسات نظرات فارسی در تجارت الکترونیک با استفاده از تابع زیان فوکال و هوش مصنوعی توضیح‌پذیر

کیانا رضائی جعفری[1]، امید مهدی عبادتی[1,*] و حمزه خواستار[2]

[1] گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران.

[2] گروه مدیریت کسب و کار، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران.

ارسال ۲۰۲۶/۰۴/۲۳؛ بازنگری ۲۰۲۶/۰۷/۰۸؛ پذیرش ۲۰۲۶/۰۸/۰۱

چکیده:

رشد سریع تجارت الکترونیک باعث افزایش حجم نظرات کاربران فارسی زبان شده است، نظراتی که حاوی اطلاعات ارزشمندی درباره محصولات و خدمات هستند. تحلیل احساسات امکان استخراج خودکار قطبیت احساسی از این داده‌ها را فراهم می‌کند. بااین‌حال، تحلیل احساسات در زبان فارسی، به‌ویژه در زمینه‌های واقعی تجارت الکترونیک، همچنان کمتر مورد توجه قرار گرفته است. علاوه بر این، عدم توازن کلاسی در مجموعه‌داده‌های تحلیل احساسات، چالش مهمی ایجاد می‌کند و اغلب به توسعه مدل‌هایی سوگیرانه منجر می‌شود که عملکرد ضعیف‌تری در تشخیص کلاس‌های اقلیت دارند. در این پژوهش، تحلیل احساسات نظرات فارسی در تجارت الکترونیک با مقایسه مدل‌های کلاسیک یادگیری ماشین، شامل رگرسیون لجستیک و ماشین بردار پشتیبان (SVM)، با مدل‌های مبتنی بر معماری ترنسفورمر، یعنی ParsBERT و ParsRoBERTa، مورد بررسی قرار گرفته است. به‌منظور کاهش تأثیر عدم توازن کلاسی، عملکرد تابع زیان فوکال در مقایسه با تابع زیان آنتروپی متقاطع مورد بررسی قرار گرفته است. همچنین، در چارچوب هوش مصنوعی توضیح‌پذیر (XAI)، از روش گرادیان‌های یکپارچه (Integrated Gradients) به‌منظور افزایش تفسیرپذیری مدل و تحلیل میزان مشارکت ویژگی‌ها استفاده شده است. نتایج آزمایش‌های انجام‌شده بر روی مجموعه‌داده دیجی‌کالا نشان می‌دهد که مدل ParsBERT آموزش‌دیده با تابع زیان فوکال، بهترین عملکرد را به دست آورده است و به دقت متوازن (Balanced Accuracy) برابر با ۸۸.۶۴ درصد و سطح زیر منحنی (AUC-ROC) ROC برابر با ۹۵.۹۱ درصد دست یافته است. یافته‌های این پژوهش نشان می‌دهند که ترکیب توابع زیان آگاه از عدم توازن کلاسی با معماری‌های مبتنی بر ترنسفورمر می‌تواند در بهبود تشخیص کلاس اقلیت در تحلیل احساسات نظرات فارسی مؤثر باشد.

کلمات کلیدی: تحلیل احساسات، پردازش زبان طبیعی فارسی، عدم توازن کلاسی، تابع زیان فوکال، هوش مصنوعی توضیح‌پذیر، نظرات تجارت الکترونیک.