Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260150
Published: June 9, 2026

A Large-Scale Comparative Study of Machine Learning, Deep Learning, and Transformer Models for Persian E-Commerce Sentiment Analysis

Roohollah Jahanmahin and Sara Masoud
Department of Industrial & Systems Engineering
Wayne State University
Detroit, MI, USA

Abstract

Sentiment analysis plays a critical role in understanding customer opinions, boosting market confidence, and guiding decision-making for businesses and governments. This study presents a comprehensive comparative analysis of sentiment classification architectures for large-scale Persian e-commerce reviews. A dataset consisting of 93,868 labeled reviews was used to evaluate traditional machine learning models (Support Vector Machine and Random Forest), deep learning models with static Word2Vec embeddings (Convolutional Neural Network and Bi-directional Long Short-Term Memory), and a transformer-based contextual model (ParsBERT). Performance was assessed using accuracy, macro precision, macro recall, and macro F1 score across varying training data proportions to examine scalability and data efficiency. Results demonstrate a clear generational progression in performance. Traditional machine learning models exhibit stable but early-saturating behavior. Static embedding-based deep learning models significantly improve scalability and balanced class performance. ParsBERT achieves the highest overall accuracy (86.47%) and macro F1 score while also demonstrating superior performance in low-data regimes, reflecting the benefits of contextual transformer-based representations. The findings highlight the importance of representation learning for sentiment analysis in morphologically rich languages such as Persian and provide practical guidance for deploying scalable sentiment monitoring systems in e-commerce environments.

Keywords
Text Mining, Deep Learning, Sentiment Analysis, Natural Language Processing.

1. Introduction
Natural language processing (NLP) is a subfield of artificial intelligence that enables computers to understand, process, and interact with human language. NLP systems convert unstructured text into machine readable representations to extract meaning and identify relationships among concepts. Recent advances in NLP have enabled applications such as voice assistants, machine translation, and grammar correction tools, demonstrating its impact across daily life, education, therapy, and industry. Sentiment analysis is a key area within NLP that focuses on identifying opinions, emotions, and attitudes toward entities such as products, services, or events (Liu, 2012; Emakhu et al., 2023).

Although sentiment analysis and opinion mining are often used interchangeably, opinions typically refer to explicit viewpoints, while sentiments reflect emotional states (Emakhu et al., 2023). The rapid growth of user generated content on platforms such as X (formerly Twitter), Meta (formerly Facebook), and Amazon has increased the importance of sentiment analysis for understanding customer feedback, attracting significant attention from both academia and industry (Liu, 2015). Deep learning has become an effective approach for NLP tasks, particularly sentiment analysis. Unlike traditional machine learning methods that rely on manual feature engineering, deep learning models automatically learn representations from data. This study designs and compares machine learning and deep learning models for Persian sentiment analysis using customer reviews from the Digi-Kala dataset collected by the

Miras Group (Rojas-Barahona and Maria, 2016). The dataset includes approximately 100,000 product reviews that are preprocessed using normalization, tokenization, and lemmatization. Five models are evaluated, including Support Vector Machines (SVM) and Random Forests (RF) with TF/IDF features, a one-dimensional convolutional neural network (CNN) with ReLU activation, and a bidirectional long short-term memory network (Bi-LSTM) using Word2Vec embeddings, and a transformer-based contextual model (ParsBERT).

Persian sentiment analysis remains challenging despite deep learning due to the language's morphological richness, flexible syntax, and widespread use of informal and colloquial expressions in user-generated text. In addition, high-quality, large-scale annotated datasets for Persian are still limited compared to high-resource languages, which constrains effective model training. These factors, combined with class imbalance and frequent spelling variations, make robust sentiment classification more difficult even with advanced deep learning models.

The primary contributions of this study are threefold. First, it provides a large-scale empirical comparison of traditional machine learning, embedding-based deep neural architectures, and contextual transformer-based models for Persian sentiment analysis using real-world e-commerce data. Second, it evaluates model robustness under class imbalance using threshold-independent metrics, highlighting performance disparities that are obscured by accuracy alone. Third, it offers practical insights into the trade-offs between computational efficiency and contextual modeling capacity in industrial deployment scenarios.

2. Literature Review

Sentiment analysis has undergone substantial methodological evolution over the past two decades, transitioning from lexicon-driven heuristics to representation learning and, more recently, large-scale pretrained language models. This progression reflects a broader shift in natural language processing (NLP) from feature engineering toward end-to-end modeling of semantic and contextual structure. To position the present study within this trajectory, the literature is reviewed across five interconnected strands: foundational sentiment classification methods, traditional supervised learning approaches, deep neural architectures, contextualized language models, and Persian-specific sentiment analysis research.

2.1 Foundational Approaches to Sentiment Classification

Early sentiment analysis research was largely grounded in lexicon-based and rule-based methods. Turney (2002) introduced an unsupervised framework that estimated semantic orientation through pointwise mutual information, while Pang et al. (2002) demonstrated the viability of supervised sentiment classification using bag-of-words features and machine learning classifiers. These foundational studies established sentiment analysis as a tractable text classification problem and showed that statistical learning methods outperformed simple polarity dictionaries.

Liu (2012, 2015) later formalized opinion mining as a structured problem involving opinion holders, targets, and sentiment expressions, thereby expanding the conceptual boundaries of sentiment analysis beyond polarity detection. Despite their interpretability and computational efficiency, lexicon-based approaches are inherently limited in modeling contextual dependencies, negation structures, and domain adaptation. These limitations become more pronounced in morphologically rich and syntactically flexible languages such as Persian.

2.2 Traditional Supervised Machine Learning Models

The second phase of sentiment analysis research emphasized supervised machine learning models trained on engineered textual features. Vector space representations such as term frequency–inverse document frequency (TF–IDF) enabled scalable document representation, while classifiers including Support Vector Machines (SVM) and ensemble methods such as Random Forest achieved competitive performance across domains.

SVMs, in particular, demonstrated strong generalization properties in high-dimensional feature spaces, making them a standard baseline for text classification. Random Forest models provided robustness against overfitting through bootstrap aggregation and feature randomness. However, both approaches rely on sparse, surface-level representations that treat documents as unordered token collections. Consequently, they fail to capture compositional semantics and long-range dependencies, limiting their effectiveness in detecting nuanced or context-dependent sentiment. These structural constraints provide a methodological rationale for incorporating deep learning architectures in sentiment classification tasks.

2.3 Deep Neural Architectures for Sentiment Modeling

The emergence of deep learning fundamentally altered NLP by enabling models to learn distributed representations directly from raw text. Convolutional Neural Networks (CNNs) were first shown to be highly effective for sentence classification tasks by Kim (2014), who demonstrated that convolutional filters can capture local semantic patterns analogous to n-grams while maintaining computational efficiency.

Recurrent neural networks (RNNs), and particularly Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), addressed the limitations of shallow models by explicitly modeling sequential dependencies. Bidirectional LSTM architectures further improved contextual sensitivity by incorporating both forward and backward representations of token sequences. In sentiment analysis, these architectures have proven especially effective in modeling negation, contrastive conjunctions, and long-distance contextual cues.

While CNNs excel at identifying localized sentiment indicators, recurrent architectures provide superior modeling of sequential and compositional semantics. Comparative evaluations across languages consistently report improved macro-level performance metrics when embedding-based deep learning models replace TF–IDF-based classifiers. The present study builds upon this literature by systematically comparing CNN and Bi-LSTM architectures against classical baselines under a consistent preprocessing and evaluation framework.

2.4 Word Embeddings and Distributed Representations

A critical development enabling deep learning in NLP was the introduction of distributed word representations. Word2Vec (Mikolov et al., 2013) demonstrated that dense vector embeddings trained on large corpora encode syntactic and semantic relationships in continuous space. Unlike sparse count-based features, embeddings preserve contextual similarity and reduce dimensionality, enhancing model generalization.

Subsequent methods such as GloVe (Pennington et al., 2014) integrated global co-occurrence statistics, further improving representation quality. For sentiment analysis, embedding-based models have been shown to outperform traditional feature engineering approaches, particularly in low-resource or morphologically complex settings. The adoption of pretrained embeddings in this study reflects this established paradigm shift and provides a theoretically grounded contrast with TF–IDF-based machine learning models.

2.5 Transformer-Based Language Models

Recent advances in NLP are dominated by the transformer architecture. Vaswani et al. (2017) introduced the self-attention mechanism, eliminating recurrence and enabling parallelized sequence modeling. Devlin et al. (2019) subsequently proposed BERT, a bidirectional transformer pretrained on large corpora, achieving state-of-the-art performance across text classification tasks.

For Persian language processing, ParsBERT (Farahani et al., 2020) extended this paradigm through large-scale pretraining on Persian corpora. Transformer-based models have demonstrated superior contextual understanding and improved performance in sentiment classification benchmarks. In addition to traditional machine learning and deep neural architectures, this study evaluates a transformer-based model (ParsBERT) to provide a comprehensive comparison across three generations of sentiment classification methods.

2.6 Persian Sentiment Analysis: Datasets and Linguistic Challenges

Persian remains comparatively underrepresented in large-scale NLP research. The language presents several challenges, including rich inflectional morphology, optional diacritics, orthographic variation between Persian and Arabic characters, flexible word order, and frequent informal usage in online contexts.

Efforts to address these challenges include the development of corpora such as MirasText (Sabeti et al., 2019) and crowdsourced Instagram sentiment datasets (Heidari and Shamsinejad, 2020). Bokaee Nezhad and Deihimi (2019) proposed a deep learning framework for Persian sentiment classification, demonstrating improvements over traditional classifiers. More recently, transformer-based approaches such as ParsBERT have further advanced Persian NLP performance.

Nevertheless, several gaps remain. Many studies rely on relatively small or domain-specific datasets, limiting generalizability. Comparative evaluations across machine learning and deep learning models are often inconsistent in

preprocessing and metric selection. Moreover, the impact of class imbalance is rarely examined in depth. By employing a large-scale e-commerce dataset and systematically comparing classical and deep learning models using multiple evaluation metrics, the present work contributes to addressing these limitations.

To synthesize the methodological progression discussed above and to clarify the positioning of the present study, Table 1 provides a structured comparison of representative sentiment analysis works across different research eras. The table highlights differences in language, dataset scale, modeling paradigm, primary strengths, and limitations, thereby making explicit the research gap addressed in this work.

Table 1. Comparative synthesis of representative sentiment analysis studies across methodological eras, highlighting architectural evolution, dataset scale, limitations, and positioning of the present study.

Era	**Representative Study**	**Language**	**Architecture**	**Scale**	**Limitation**	**Relation to This Work**
Classical ML	Pang et al. (2002)	English	SVM + BoW	Small	Sparse features	Baseline lineage
Representation Learning	Kim (2014)	English	CNN	Benchmark	Local focus	CNN grounding
Sequential Modeling	Hochreiter & Schmidhuber (1997)	General	LSTM	—	No bidirectionality originally	Bi-LSTM rationale
Persian DL	Bokaee Nezhad & Deihimi (2019)	Persian	DL	Small	Limited scale	Persian precedent
Transformer Era	Farahani et al. (2020)	Persian	ParsBERT	Large	Heavy computation	Efficient alternative
This Study	—	Persian	ML + CNN + Bi-LSTM	Large-scale real reviews	No transformer	Comparative evaluation

As shown in Table 1, sentiment analysis research has progressed from sparse feature–based classifiers to representation learning and, more recently, transformer-based contextual models. Early supervised approaches such as Pang et al. (2002) established sentiment classification as a viable machine learning problem but were limited by bag-of-words representations that ignored contextual dependencies. The introduction of convolutional architectures (Kim, 2014) enabled local semantic feature extraction, while recurrent networks such as LSTMs addressed sequential modeling and long-range contextual interactions. In the Persian language domain, deep learning adoption has demonstrated clear improvements over traditional classifiers; however, most studies have relied on relatively small datasets or domain-constrained corpora.

More recent transformer-based approaches, including ParsBERT (Farahani et al., 2020), leverage large-scale pretraining to achieve state-of-the-art performance in Persian NLP tasks. Nevertheless, such models introduce substantial computational complexity and are not always practical for industrial deployment scenarios. The present study contributes to this landscape by providing a systematic comparison of traditional machine learning and deep learning architectures on a large-scale real-world Persian e-commerce dataset. By evaluating multiple metrics and explicitly analyzing class imbalance effects, this work addresses gaps in scale, methodological consistency, and practical interpretability that remain underexplored in prior Persian sentiment analysis research.

3. Methods

This study performs sentence level sentiment analysis on a Persian online shopping dataset. Sentence level analysis is adopted to limit feature dimensionality and maintain classification accuracy. Figure 1 illustrates the overall methodology adopted in this study. The pipeline begins with a labeled dataset of Persian user comments, which undergoes comprehensive text preprocessing, including punctuation removal, tokenization, stop word elimination, stemming, lemmatization, and normalization. The processed text is then represented using two different feature extraction strategies: TF–IDF for traditional machine learning models and Word2Vec embeddings for CNN and Bi-LSTM models and also BERT tokenizer for ParsBERT model. Sentiment classification is performed using Support

Vector Machine (SVM) and Random Forest (RF) classifiers, as well as CNN and Bi-LSTM architectures, and a ParsBERT model. Model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and ROC–AUC, to provide a comprehensive assessment of classification effectiveness under class imbalance.

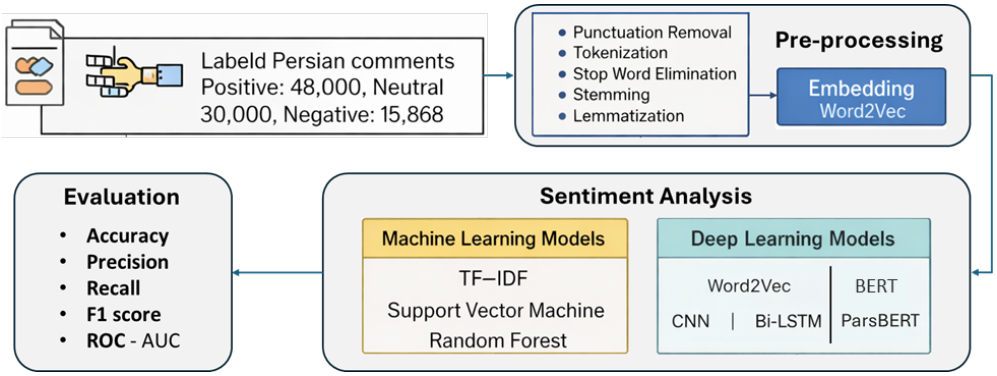

Figure 1. End-to-end workflow of the proposed Persian sentiment analysis methodology

3.1. Pre-processing

Text preprocessing comprises punctuation removal, tokenization, stop-word elimination, stemming, lemmatization, and normalization. Punctuation marks are removed as they introduce structural noise without contributing semantic information. Tokenization segments the text into individual lexical units, enabling subsequent analysis. Common Persian stop words, such as "از", "به", and "زیرا", are eliminated to reduce dimensionality and model complexity. Given the rich morphological structure of the Persian language, stemming and lemmatization are performed using a language-specific lexical analyzer rather than English-based algorithms. Text normalization standardizes equivalent Persian and Arabic characters into a unified form, ensuring consistency across the corpus. Following preprocessing, textual data are transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF) and Word2Vec embeddings. TF-IDF assigns weights to terms based on their frequency within a document and their inverse frequency across the corpus, thereby highlighting discriminative features for classification. Word2Vec learns dense vector representations by modeling contextual word co-occurrence, preserving semantic relationships and enabling effective representation learning for deep learning-based sentiment analysis models.

3.2. Machine Learning Models

Sentiment classification is formulated as a supervised learning problem. Support Vector Machines are trained to find an optimal separating hyperplane for sentiment classes, defined as $H = \{x \mid w^T x + b = 0\}$. Random Forest is employed as an ensemble method in which each decision tree is trained on a random subset of features. Given base learners $h_1(x), \ldots, h_j(x)$, final predictions are obtained via majority voting, improving robustness and reducing overfitting. Hyperparameter tuning is performed using grid search with 5-folds cross-validation. For Random Forest, the number of estimators (e.g., 100–500), maximum depth (e.g., 10), minimum samples split (e.g., 2–10), and minimum samples per leaf (e.g., 1–4) are explored. For SVM, the regularization parameter C (e.g., 0.1–10), kernel type (linear, RBF), and kernel coefficient (gamma) are tuned.

3.3. Deep Learning Models

3.3.1. Convolutional Neural Network (CNN)

A one-dimensional CNN is employed for sentence-level sentiment classification. Input sentences are tokenized and converted into fixed-length sequences through padding or truncation. Each token is mapped to a dense vector using a pre-trained Word2Vec embedding matrix. The embedding layer is initialized with these weights and set to be trainable, enabling task-specific fine-tuning during supervised learning. As displayed in Figure 2, the CNN architecture comprises three convolutional layers designed to capture local textual patterns at multiple n-gram scales. Each layer uses 64 filters with kernel sizes of 4, 8, and 16, respectively, ReLU activation, and same padding. Max pooling with a pool size of 2 is applied after the first two convolutional layers to progressively reduce sequence length while retaining salient features. The final convolutional layer is followed by a global max pooling operation to aggregate the most informative features across the entire sequence. The extracted features are passed through a dropout layer to mitigate overfitting, followed by a fully connected dense layer with 500 neurons and ReLU activation for high-level representation learning. The output layer consists of three neurons with a softmax activation function corresponding

to the sentiment classes. The model is trained using categorical cross-entropy loss and the Adam optimizer with a batch size of 64 for 10 epochs. Due to computational constraints hyperparameters, including the learning rate (0.001 and 0.002) and dropout rate (0.10 and 0.15), are selected via grid search with three-fold cross-validation on the training set. The best-performing configuration is subsequently evaluated on a held-out test set.

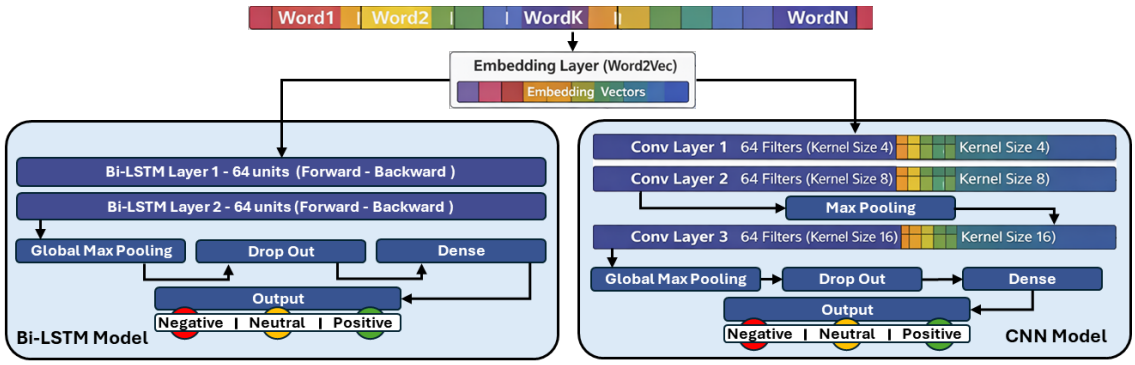

Figure 2. Architectural overview of the proposed deep learning models for Persian sentiment analysis, illustrating the Word2Vec embedding layer followed by a stacked Bi-LSTM network and a multi-kernel CNN

3.3.2. Bidirectional Long Short-Term Memory (Bi-LSTM)

A Bi-LSTM network is used to model sequential dependencies by processing token sequences in both forward and backward directions. As in the CNN model, input sentences are padded or truncated to a fixed length, and a pre-trained Word2Vec embedding matrix is used to initialize a trainable embedding layer. The Bi-LSTM architecture (as illustrated in Figure 2) consists of two stacked bidirectional LSTM layers, each with 64 hidden units and configured with return_sequences = True to preserve time-step representations. The resulting sequence output is reduced using a GlobalMaxPooling1D layer, followed by dropout regularization and a dense layer with 500 neurons and ReLU activation. The final output layer uses a three-unit softmax activation to produce class probabilities for the sentiment categories. The Bi-LSTM model follows the same training and validation protocol as the CNN, including the Adam optimizer, categorical cross-entropy loss, batch size of 64, 10 training epochs, and three-fold cross-validation for hyperparameter selection and model evaluation.

3.3.3. Transformer-Based Model: ParsBERT

ParsBERT is a recently developed model that is less resource-intensive compared to multilingual BERT (Farahani et al. 2021). It demonstrates cutting-edge performance in various practical applications, including Sentiment Analysis, Text Classification, and Named Entity Recognition (Farahani et al. 2021). To improve the generality and efficiency of this pre-trained model at different linguistic levels, (Farahani et al. 2021) created a new corpus by collecting data from diverse sources such as Persian Wikipedia**,** BigBangPage**,** Chetor**,** Eligasht**,** Digikala**,** Ted Talks subtitles**,** fictional books**,** novels**,** and MirasText, which extensively crawled over 250 Persian news websites. After collecting the pre-training corpus, it undergoes a crucial two-step processing, involving tasks like cleaning, replacing, sanitizing, and normalizing to ensure the dataset is properly formatted (Farahani et al. 2021).

The ParsBERT model is structured on the BERT model architecture, featuring 12 hidden layers, 12 attention heads, and a hidden size of 768, totaling 110 million parameters. The pre-training involves two tasks: 1) a Masked Language Model (MLM), where the model predicts randomly masked tokens, and 2) a Next Sentence Prediction (NSP) task, predicting if the second sentence in a pair follows the first (Farahani et al. 2021).

3.4. Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, and F1 score to capture complementary aspects of classification quality. Receiver Operating Characteristic curves and Area Under the Curve values are also reported to assess discriminative ability, particularly under class imbalance. Higher AUC values indicate stronger separation between sentiment classes and more reliable generalization to unseen data.

4. Results and Discussion

This section evaluates the performance of machine learning and deep learning models for multi-class sentiment analysis, focusing on the classification of user comments into positive, neutral, and negative categories. The dataset

analyzed in this study was compiled by the Miras Group and consists of user opinions collected from Digi-Kala, Iran's largest online shopping platform. The dataset includes a total of 93,868 unique comments, where the first column contains the textual user comments and the second column provides the corresponding sentiment labels (Rojas-Barahona and Maria, 2016). Among these comments, 48,000 are labeled as positive, 30,000 as neutral, and 15,868 as negative, as illustrated in Figure 3-a. This distribution reveals a notable class imbalance, particularly for the negative sentiment category, which has implications for both model training and evaluation.

Figure 3-b presents the distribution of comment lengths in the dataset. Most entries are short to moderately long, with approximately 40% of comments containing fewer than 18 words and nearly 90% containing fewer than 90 words. While concise comments dominate the dataset, longer entries introduce additional linguistic variability and contextual complexity, posing challenges for models that rely on fixed-length feature representations.

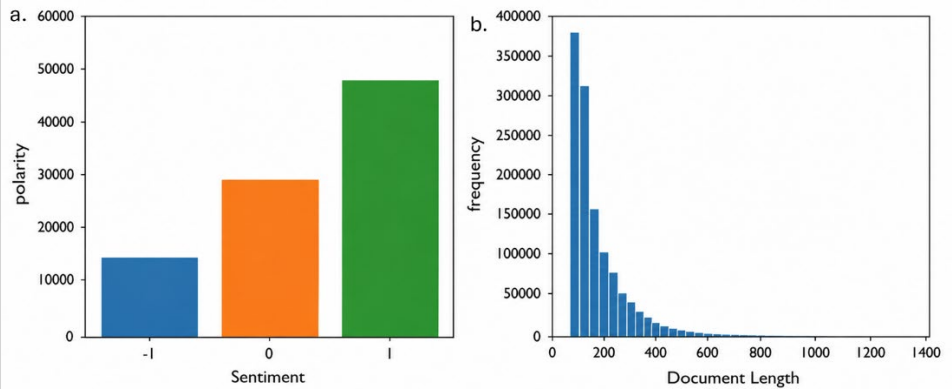

Figure 3. a) frequency chart of different categories of comments and b) histogram of sentences' length (words)

Figure 4 presents the overall accuracy comparison across traditional machine learning and deep learning architectures. ParsBERT achieves the highest performance (86.47%), followed by Bi-LSTM (80.91%) and CNN (78.79%), while traditional classifiers (SVM and RF) exhibit lower asymptotic accuracy.

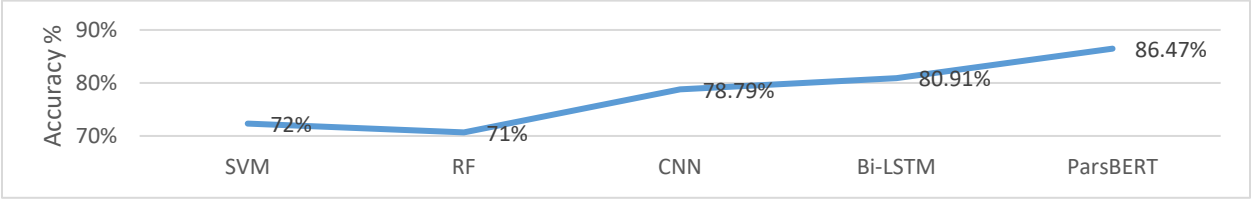

Figure 4. Overall accuracy of the evaluated models, including SVM, RF, CNN, Bi-LSTM, and ParsBERT

4.1. Traditional Machine Learning Models

Beyond overall accuracy, model stability and class-level discrimination are analyzed using performance–precision distributions and ROC curves.

4.1.1. Performance Stability Across Training Sizes

Figures 5(a) and 5(c) present the joint distribution of accuracy and precision for SVM and RF across increasing training data proportions. Each point corresponds to a different percentage of the available labeled data, enabling assessment of scalability and stability. The SVM classifier demonstrates relatively stable performance across training sizes, with accuracy fluctuating within a narrow band (approximately 0.78–0.81). Precision values remain similarly constrained between 0.79 and 0.83. Notably, performance gains beyond 43% training size are marginal, indicating early representational saturation. While slight improvements are observed at full training scale, the incremental benefit of additional labeled data remains

limited. This behavior suggests that SVM, operating on TF-IDF representations, reaches its effective capacity relatively early and does not substantially benefit from larger datasets.

RF exhibits a comparable but slightly lower performance profile. Accuracy varies between approximately 0.75 and 0.78 across training sizes, and precision ranges from 0.78 to 0.81. Unlike deep learning models, neither SVM nor RF demonstrates a clear monotonic improvement pattern with increasing data. Instead, performance fluctuations appear modest and non-progressive, reinforcing the interpretation that classical feature-based models have limited scalability under expanding supervision. Comparatively, SVM maintains slightly higher asymptotic accuracy and tighter performance dispersion than RF, suggesting stronger margin-based separation in high-dimensional feature space.

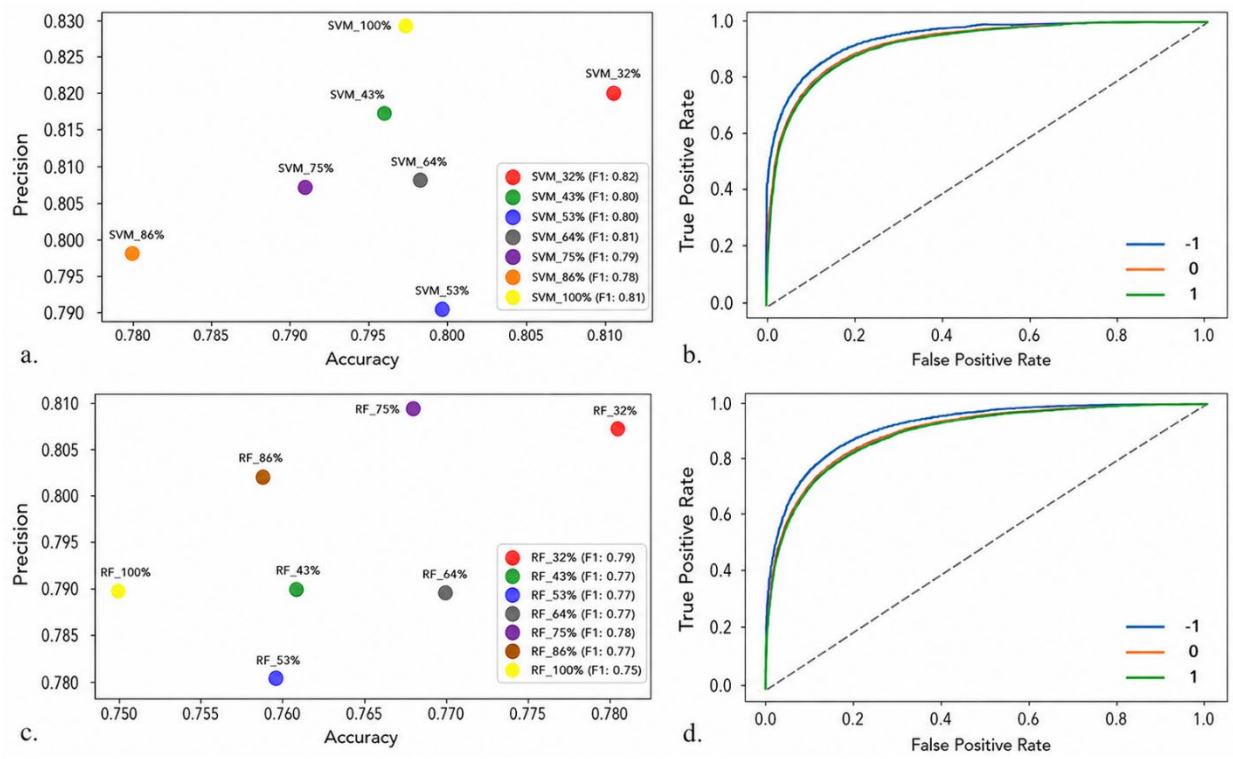

Figure 5. ROC curves of the evaluated models, including a) SVM, b) RF, c) CNN, and d) Bi-LSTM.

4.1.2. Class-Level Discrimination Analysis

Figures 5(b) and 5(d) display the multi-class ROC curves for SVM and RF, respectively. The curves lie substantially above the diagonal baseline, confirming meaningful discriminatory power across all sentiment labels (-1, 0, 1). For SVM, the AUC values for each class are approximately 0.92–0.93, indicating strong probabilistic separability despite moderate overall accuracy. Importantly, the ROC curves for negative, neutral, and positive classes closely overlap, suggesting balanced discrimination and limited bias toward any specific class. This behavior is particularly relevant in the presence of class imbalance, as it indicates that SVM preserves class ranking integrity even when threshold-based accuracy gains plateau.

RF achieves slightly lower but still robust AUC values (approximately 0.90–0.91 across classes). Similar to SVM, the ROC curves demonstrate consistent class-level separability; however, the marginal reduction in AUC aligns with the slightly lower asymptotic accuracy observed in the performance distribution analysis. As displayed in Figure 5, the ROC results reveal that while traditional machine learning models provide stable and well-calibrated probability estimates, their representational capacity limits further performance gains as dataset size increases. In contrast to deep learning architectures, which learn hierarchical feature abstractions, SVM and RF rely on static feature representations and thus exhibit early saturation behavior.

4.2. Deep Learning Models

Here, we evaluate the performance and scalability of deep learning architectures, including CNN with Word2Vec embeddings, Bi-LSTM with Word2Vec embeddings, and the transformer-based ParsBERT model. Unlike traditional classifiers, these models learn hierarchical feature representations and exhibit distinct learning behaviors as the training dataset increases. To assess scalability and balanced performance, learning curves for accuracy, macro precision, macro recall, and macro F1 score are analyzed across varying training proportions (21%–100%). Training data proportions are selected to correspond to approximately equal increments in sample size (i.e., multiples of 10,000 samples). As a result, slight deviations from standard percentage intervals (e.g., 21% instead of 20%) were introduced to maintain integer dataset sizes and ensure consistent scaling across experiments.

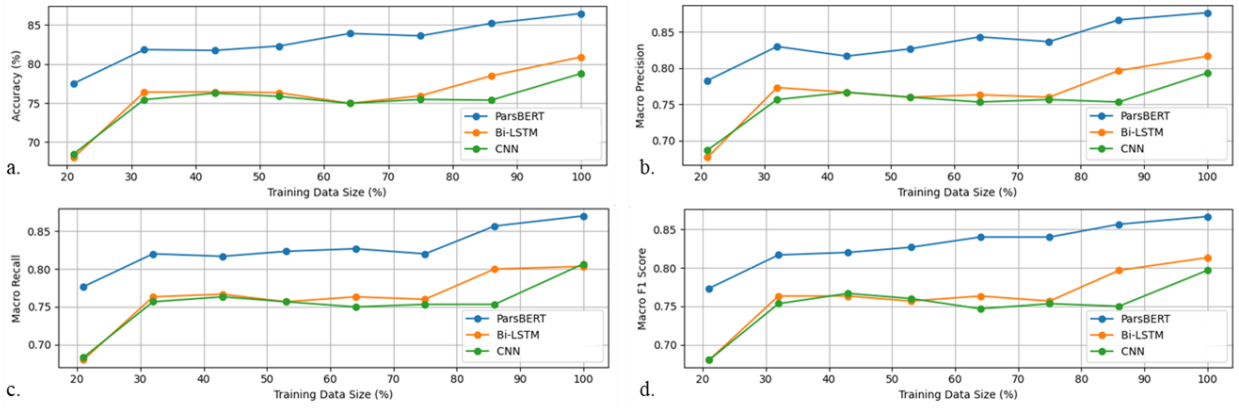

Figure 6. Learning Curves of CNN, Bi-LSTM, and ParsBERT on a) Accuracy, b) Precision, c) Recall, and d) F-1 score

4.2.1 Accuracy and Data Efficiency

Figure 6(a) presents the learning curves for classification accuracy. Clear architectural differences emerge across training regimes. ParsBERT consistently achieves the highest accuracy at all training sizes. At 21% of the available data, ParsBERT reaches 77.51%, substantially outperforming both Bi-LSTM (68.02%) and CNN (68.49%). This large performance gap in low-data conditions reflects the advantage of contextual pretraining. Because ParsBERT leverages transformer-based representations pretrained on large Persian corpora, it requires fewer labeled samples to achieve strong performance. As training size increases, all models improve; however, their growth patterns differ. Bi-LSTM demonstrates the steepest relative improvement, increasing by approximately 13 percentage points from 21% to 100% training size, ultimately reaching 80.91%. CNN exhibits a similar but slightly flatter growth trajectory, reaching 78.79% at full scale. In contrast, ParsBERT improves more gradually (approximately 9 percentage points overall), suggesting that its pretrained contextual representations already capture much of the linguistic structure required for the task. Notably, performance gains diminish beyond approximately 75% training size for all models, indicating convergence toward asymptotic capacity limits. Table 2 summarizes the classification accuracy across different training sizes for CNN, Bi-LSTM, and ParsBERT models.

Table 2. Accuracy (%) across training sizes on Deep Learning models

Training Size (%)	CNN	Bi-LSTM	ParsBERT
21%	68.49	68.02	77.51
30%	73.5	72.8	81.0
43%	74.0	74.5	82.0
57%	74.5	76.5	83.0
75%	76.5	78.5	84.5
100%	78.79	80.91	**86.47**

4.2.2 Metric-Level Performance Under Class Imbalance

Figures 6(b)–6(d) present macro precision, recall, and F1 learning curves. The use of macro-averaged metrics ensures that performance improvements reflect balanced gains across sentiment classes rather than dominance of majority categories. Across all three metrics, ParsBERT maintains superior performance at every training size. Macro F1 increases steadily from approximately 0.77 at 21% training size to approximately 0.87 at full scale. The consistency across precision and recall suggests balanced discrimination among negative, neutral, and positive labels. Bi-LSTM demonstrates strong scalability in recall, particularly at higher training proportions, indicating improved contextual sensitivity as more labeled samples become available. Macro F1 for Bi-LSTM increases from approximately 0.68 at 21% to approximately 0.81 at 100%. This confirms that recurrent architectures benefit substantially from additional supervision, narrowing, but not eliminating, the performance gap relative to transformer-based models. CNN exhibits earlier performance saturation. Although macro F1 improves from approximately 0.68 to approximately 0.80, its recall and precision curves flatten earlier than those of Bi-LSTM. This behavior is consistent with convolutional architectures' emphasis on local n-gram feature extraction rather than long-range sequential modeling. Importantly, the similarity between accuracy and macro-F1 trajectories indicate that improvements are not solely driven by dominant classes but reflect balanced class-level performance gains.

4.2.3 Architectural Interpretation

The observed learning behaviors reflect fundamental differences in representation mechanisms. CNN and Bi-LSTM rely on static Word2Vec embeddings, where each word is assigned a fixed vector representation independent of context. Consequently, these models must learn contextual relationships primarily through supervised training, resulting in stronger dependence on labeled data. ParsBERT, in contrast, employs contextual transformer-based embeddings. Token representations dynamically adapt based on surrounding words, enabling richer semantic encoding prior to task-specific fine-tuning. This contextualization explains both the strong performance in low-data regimes and the higher asymptotic accuracy observed at full training scale. While recurrent architecture demonstrates competitive scalability when large annotated datasets are available, transformer-based models provide both superior data efficiency and higher performance ceilings. However, these gains must be interpreted alongside computational complexity considerations, which are discussed in Section 4.3.

4.3 Architectural Insights and Practical Implications

A central finding of this study is the systematic performance progression across architectural paradigms, from convolutional and recurrent models to transformer-based contextual modeling. While both CNN and Bi-LSTM architectures achieve competitive performance, their learning behaviors and representational capacities differ substantially. CNN models primarily capture local n-gram–level patterns through convolutional filters, making them effective at identifying salient sentiment cues within short spans of text. This local feature extraction mechanism enables efficient learning of frequent sentiment expressions but limits sensitivity to long-range dependencies and compositional structure. As a result, CNN performance tends to saturate earlier as training data increase. Bi-LSTM models, in contrast, explicitly model sequential dependencies in both forward and backward directions. This bidirectional structure allows them to capture contextual interactions across tokens, including negation, intensifiers, and multi-word sentiment expressions. In Persian text, where sentiment meaning may depend heavily on word order and context spanning multiple tokens, this sequential modeling capability yields more balanced improvements across positive, neutral, and negative classes. Consequently, Bi-LSTM exhibits stronger scalability with increasing training size compared to CNN.

ParsBERT extends this progression further by incorporating transformer-based contextual embeddings. Unlike static Word2Vec representations used in CNN and Bi-LSTM models, ParsBERT dynamically adjusts token representations based on full-sentence context through self-attention mechanisms. This enables richer semantic encoding and superior disambiguation of nuanced sentiment expressions. The learning curve results demonstrate that ParsBERT not only achieves the highest asymptotic accuracy but also exhibits strong data efficiency in low-resource regimes. These findings confirm the advantages of contextual representation learning for sentiment classification in morphologically rich and context-sensitive languages such as Persian.

To further quantify the trade-offs between computational efficiency and contextual modeling capacity, we analyze the training and inference costs of the evaluated models. CNN demonstrates the highest computational efficiency, requiring less than one minute per cross-validation fit on smaller datasets (20k samples) and approximately 80 minutes at larger scale (80k samples), with fast inference times (~10–111 ms per step). Bi-LSTM improves sequential modeling performance but introduces significantly higher computational cost, with training times ranging from

approximately 6 minutes to over 4 hours depending on dataset size, and slower inference (~631 ms to 4 s per step). In contrast, ParsBERT provides the most expressive contextual representations but at the highest computational cost, with total training time increasing from approximately 1 hour (20k samples) to nearly 9 hours (50k samples) due to its large parameter size (~110M) and transformer-based architecture. While ParsBERT achieves the highest predictive performance, its transformer-based architecture introduces significantly higher computational cost compared to SVM, RF, CNN, and Bi-LSTM. Fine-tuning requires GPU acceleration and increased memory consumption due to multi-head self-attention layers and large parameter count (~110M parameters). In contrast, SVM and RF models train rapidly on CPU environments, making them suitable for low-resource deployment scenarios. CNN and Bi-LSTM architectures provide a middle ground, offering improved predictive power relative to traditional models with moderate computational requirements. These trade-offs are relevant in industrial settings where latency, scalability, and infrastructure constraints influence model selection.

From an applied perspective, these findings carry direct implications for large-scale e-commerce platforms. Accurate and robust sentiment analysis enables systems such as Digi-Kala to automatically monitor customer satisfaction, detect recurring product or service issues, and support data-driven operational decisions. Models that maintain balanced performance across minority sentiment classes, particularly negative reviews, are especially valuable for early detection of dissatisfaction, quality control, and proactive service intervention. The results indicate that deep learning architectures, particularly transformer-based models, substantially outperform traditional machine learning pipelines in both scalability and balanced classification performance. While CNN and Bi-LSTM architectures offer meaningful improvements over sparse feature-based methods, ParsBERT provides the strongest overall performance and greater robustness under limited data conditions. Therefore, for production-level deployment in large-scale feedback monitoring systems, contextual transformer-based architectures represent the most effective solution when computational resources permit.

5. Conclusion

This study provides a systematic evaluation of sentiment classification architectures on a large-scale Persian e-commerce dataset. By comparing traditional machine learning models, static embedding-based deep learning architectures, and a contextual transformer-based model, the results reveal distinct performance and scalability patterns. Traditional classifiers such as SVM and RF demonstrate stable and well-calibrated class discrimination but exhibit limited performance gains as training data increase. Deep learning models incorporating static Word2Vec embeddings significantly improve scalability, with Bi-LSTM outperforming CNN due to its ability to model sequential dependencies and contextual interactions. ParsBERT achieves the highest overall performance and demonstrates strong data efficiency, maintaining superior accuracy and macro-F1 scores even at reduced training sizes. These findings confirm that contextual transformer-based representations provide both improved asymptotic performance and greater robustness under varying data availability. From an applied perspective, accurate sentiment classification enables e-commerce platforms to monitor customer satisfaction, detect emerging service issues, and support data-driven quality control. Models with balanced performance across sentiment classes are critical for early detection of negative feedback. While traditional models offer computational efficiency, transformer-based architecture provides the strongest predictive capability when computational resources permit. Future work may explore lightweight transformer adaptations, domain-specific pretraining, and cost-performance trade-offs to further optimize large-scale deployment in industrial environments.

References

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 4171–4186, Minneapolis, USA, June 2–7, 2019.

Emakhu, J., Etu, E.E., Monplaisir, L., Aguwa, C., Arslanturk, S., Masoud, S., Tenebe, I.T., Nassereddine, H., Hamam, M. and Miller, J., A hybrid machine learning and natural language processing model for early detection of acute coronary syndrome, Healthcare Analytics, vol. 4, p. 100249, 2023.

Farahani, M., Gharachorloo, M., Farahani, M. and Manthouri, M., ParsBERT: Transformer-Based Model for Persian Language Understanding, arXiv preprint arXiv:2005.12515, 2020.

Heidari, M. and Shamsinejad, P., Producing an Instagram dataset for Persian language sentiment analysis using crowdsourcing method, Proceedings of the 6th International Conference on Web Research (ICWR), pp. 284–287, Tehran, Iran, April 22–23, 2020.

Hochreiter, S. and Schmidhuber, J., Long Short-Term Memory, Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

Kim, Y., Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, Doha, Qatar, October 25–29, 2014.

Liu, B., Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.

Liu, B., Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., Efficient estimation of word representations in vector space, Proceedings of the International Conference on Learning Representations (ICLR) Workshop, Scottsdale, USA, May 2–4, 2013.

Miras-Tech, MirasText, Available: https://github.com/miras-tech/MirasText
, Accessed 2024.

Mishra, K., Kandasamy, I., Vasantha Kandasamy, W.B. and Smarandache, F., A novel framework using neutrosophy for integrated speech and text sentiment analysis, Symmetry, vol. 12, no. 1, pp. 1–22, 2020.

Onyenwe, I., Nwagbo, S., Mbeledogu, N. and Onyedinma, E., The impact of political party/candidate on the election results from a sentiment analysis perspective using #AnambraDecides2017 tweets, Social Network Analysis and Mining, vol. 10, 2020.

Oueslati, O., Cambria, E., HajHmida, M.B. and Ounelli, H., A review of sentiment analysis research in Arabic language, Future Generation Computer Systems, vol. 112, pp. 408–430, 2020.

Pang, B., Lee, L. and Vaithyanathan, S., Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, Philadelphia, USA, July 6–7, 2002.

Pennington, J., Socher, R. and Manning, C.D., GloVe: Global vectors for word representation, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, Doha, Qatar, October 25–29, 2014.

Rojas-Barahona, L. and Maria, L., Deep learning for sentiment analysis, Language and Linguistics Compass, vol. 10, no. 12, pp. 1–16, 2016.

Roshanfekr, B., Khadivi, S. and Rahmati, M., Sentiment analysis using deep learning on Persian texts, Proceedings of the 25th Iranian Conference on Electrical Engineering (ICEE), pp. 1503–1508, Tehran, Iran, May 2–4, 2017.

Sabeti, B., Abedi Firouzjaee, H., Janalizadeh Choobbasti, A., Mortazavi Najafabadi, S.H.E. and Vaheb, A., Mirastext: An automatically generated text corpus for Persian, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pp. 1174–1177, Miyazaki, Japan, May 7–12, 2018.

Turney, P.D., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417–424, Philadelphia, USA, July 6–12, 2002.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., Attention is all you need, Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, Long Beach, USA, December 4–9, 2017.

Zuheros, C., Martínez-Cámara, E., Herrera-Viedma, E. and Herrera, F., Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid: Case study of restaurant choice using TripAdvisor reviews, Information Fusion, vol. 55, pp. 22–36, 2020.

Biographies

Roohollah Jahanmahin received his B.S. in Industrial Engineering from University of Tehran in 2018 and his M.S. in Industrial Engineering, with a concentration in Optimization Systems, from Iran University of Science and Technology in 2021. He is currently a Ph.D. Candidate Research Assistant in the Department of Industrial and Systems Engineering at Wayne State University and is affiliated with the Smart Immersive Modeling Lab. His research lies at the intersection of human–vehicle interaction and autonomous driving, with a strong emphasis on immersive Virtual Reality (VR) simulation. His work focuses on developing human-centered, data-driven frameworks to understand and predict driver behavior, enabling personalized control transitions, and enhancing the safety and adaptability of semi-autonomous systems in mixed-traffic environments.

Dr. Sara Masoud is an Assistant Professor of Industrial and Systems Engineering at Wayne State University whose research centers on human-centered engineering, immersive technologies, and data-driven modeling of complex energy and infrastructure systems. She integrates advanced analytics, simulation, and virtual, augmented, and mixed reality environments to enhance operator training, workforce development, and system resilience in high-risk domains.

Her work examines how human cognition, trust, and decision-making affect system performance under both normal and abnormal conditions. Dr. Masoud leads multidisciplinary, externally funded projects across energy, manufacturing, transportation, and smart agriculture, producing methodological frameworks and immersive platforms that connect research with real-world deployment.