Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Fusion of Statistical and Stylistic Text Features with SVM for Persian Sentiment Analysis

Alireza Bahmani *

Department of Information Technology, Islamic Azad University Central Tehran Branch, Tehran 1955847781, Iran; e-mail : alirezabahmani88@gmail.com

* Corresponding Author : Alireza Bahmani

Abstract: Sentiment analysis is a critical task in natural language processing (NLP) that classifies text into sentiment categories, such as positive, negative, or neutral. This task is particularly challenging for languages like Persian due to the complexity of their linguistic structure and the scarcity of high-quality labeled datasets. Previous studies on Persian sentiment analysis have largely relied on TF-IDF representations or deep learning models, often overlooking handcrafted statistical and stylistic features that capture subtle textual patterns. This limitation reduces their effectiveness, especially when dealing with informal or noisy text data. Experiments in this study were conducted on a dataset of Persian product reviews from Digikala.com, labeled according to user ratings to indicate positive, negative, or neutral sentiment. In this paper, we propose a novel approach to Persian text sentiment analysis by combining statistical and stylistic (surface-level) features with traditional text-based features such as Term Frequency–Inverse Document Frequency (TF-IDF). Unlike prior works that rely solely on TF-IDF or deep learning representations, our method integrates stylistic and statistical cues to capture expressive nuances in informal Persian text. Additionally, the Support Vector Machine (SVM) classifier is optimized using RandomizedSearchCV to enhance performance. The proposed system utilizes both statistical and textual features to improve classification accuracy. We compare its performance with four baseline models, i.e., Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree, that rely solely on TF-IDF features. The experimental results demonstrate that the proposed approach outperforms the baseline models in terms of accuracy, F1-score, recall, and precision. Specifically, the proposed system achieved the highest accuracy (0.8354), significantly improving negative sentiment detection while maintaining strong performance in positive sentiment classification.

Keywords: Feature Fusion; Machine Learning; Persian Language Processing; Sentiment Analysis; Statistical Features; Stylistic Features; Support Vector Machine (SVM); TF-IDF Representation.

Received: October, 18th 2025
Revised: November, 27th 2025
Accepted: November, 29th 2025
Published: December, 5th 2025
Curr. Ver.: December, 5th 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

In today's world, sentiment analysis has become one of the most important challenges in Natural Language Processing (NLP). This field of data processing has attracted significant attention, particularly across various languages. Sentiment analysis generally refers to the identification and extraction of users' opinions and emotions from different texts, which can be categorized into positive or negative sentiments, or even more complex categories such as happy, sad, or angry [1]. These analyses have widespread applications in customer feedback analysis, social media monitoring, surveys, and in cases where public sentiment and opinions are crucial [2]. At the same time, recent systematic reviews have documented significant progress in state-of-the-art sentiment analysis methods, reflecting the field's growing maturity and methodological breadth [3].

Among the various languages, Persian is considered one of the most complex in text processing due to its unique structural and grammatical features. As a result, sentiment analysis in this language requires more specific and accurate approaches [4]. In Persian, the earliest attempts used lexicon-driven approaches to address language-specific challenges, such as orthographic variation and morphological complexity, to extract sentiment from text [5]–[7].

The problem of sentiment analysis in Persian, especially in the context of user reviews and textual feedback, is significant because it enables effective analysis of large datasets and extraction of valuable insights from public opinions and viewpoints on Persian-language platforms. Due to its distinctive grammatical and structural characteristics, Persian differs considerably from Western languages such as English. Some of these features include compound verbs, informal words and expressions, phonetic variations, non-standard punctuation, and specialized terminology, all of which require more precise methods and complex algorithms for proper analysis. Therefore, the development of tools that accurately identify sentiment in Persian texts and correctly classify them is essential.

With recent advancements in machine learning and artificial intelligence, many of these issues have been effectively addressed [8]. Machine learning, particularly Support Vector Machine (SVM)-based classification methods, has emerged as a powerful approach for analyzing text. These algorithms have demonstrated significant performance, especially when dealing with large amounts of textual data. However, to improve the accuracy of such models and address the challenges specific to Persian, it is necessary to leverage more complex textual features and adopt hybrid approaches.

Previous studies on Persian sentiment analysis have largely relied on traditional TF-IDF representations or deep learning models. While TF-IDF-based approaches are computationally efficient, they often fail to capture informal linguistic patterns, stylistic cues, and subtle expressive nuances in user-generated texts. Deep learning models such as BERT, BiLSTM, or CNN can address some of these limitations but typically require large labeled datasets and substantial computational resources, which are often unavailable for Persian. Therefore, hybrid approaches that integrate TF-IDF with surface-level statistical and stylistic features are required to achieve a balance between performance, efficiency, and the ability to capture nuanced sentiment expressions in Persian text.

This study explores how machine learning and natural language processing techniques can be applied to enhance sentiment analysis for Persian texts. The research aims to introduce methods that leverage linguistic and structural features of Persian text to achieve more accurate sentiment classification of users' opinions in online reviews and other types of feedback. In this context, employing hybrid techniques that utilize both statistical and linguistic information from Persian texts can significantly improve the performance of sentiment analysis models.

Despite the growing number of studies on Persian sentiment analysis, many existing approaches rely primarily on TF-IDF or word-level embeddings and often overlook informal writing patterns commonly found in user-generated reviews. Existing models also show limited capability in detecting negative sentiment, partly due to class imbalance and the absence of complementary handcrafted features. Furthermore, few studies have systematically combined statistical cues such as repeated characters, digit ratios, or emphasis markers with textual features, leaving a gap in capturing informal emotional expressions in Persian product reviews.

Support Vector Machines (SVMs) were selected because they are highly effective for high-dimensional, sparse representations such as TF-IDF, require relatively small datasets compared to deep learning models, and deliver strong performance with limited computational resources. Since the feature space in this study combines sparse textual features with handcrafted statistical ones, SVM offers a robust and efficient choice that avoids the large data requirements and computational overhead associated with deep neural models.

The main contributions of this study are summarized as follows:

- A feature-enhanced sentiment analysis model is proposed, which combines TF-IDF representations with handcrafted statistical features tailored for informal Persian reviews.
- An optimized SVM-based classification framework is designed, and an ablation study is conducted to quantify the contribution of each feature domain.
- The proposed method is evaluated on a real-world Persian product review dataset, demonstrating improvements over traditional baseline models.

The remainder of this paper is organized as follows: Section 2 presents the related work; Section 3 describes the proposed methodology; Section 4 discusses the experimental results and analysis; and Section 5 concludes the paper with key findings and suggestions for future research.

2. Related Work

The study [9] developed a novel Persian sentiment analysis architecture using deep learning techniques to classify movie and hotel reviews. Shallow and deep learning algorithms were compared, and it was found that a stacked bidirectional LSTM achieved the highest accuracy of 95.61% for movie reviews, while the 2D-CNN model outperformed others for hotel reviews with an accuracy of 89.76%. Another study [10] presented a context-aware multimodal sentiment analysis framework that utilized textual, acoustic, and visual cues. A Persian multimodal dataset with over 800 utterances was introduced, and CNN and LSTM models were applied for automated feature extraction. The results showed that the multimodal approach, combining all feature modalities, outperformed unimodal methods, achieving 91.39% accuracy compared to 89.24%.

A recent investigation [11] focused on sentiment analysis of Persian user reviews from the Iranian online store Digikala. It addressed the challenges of unstructured Persian text and the lack of preprocessing tools by applying web-mining techniques to gather approximately three million reviews. The fastText method was used for word embedding to reduce preprocessing time and was compared with TF-IDF. Using CNN and fastText, the approach achieved an AUC of 0.996 and an F-score of 0.956, outperforming models such as BiLSTM and Logistic Regression. Research by [12] explored the detection of abusive language in Persian tweets, a growing concern amid the rise of online interactions. Due to the lack of suitable datasets, a collection of 33,338 Persian tweets was created, including 10% abusive content. A simple list-based approach achieved 76% accuracy, while a deep learning model using BERT achieved 97.7%.

An approach proposed in [13] introduced a deep learning-based architecture for Persian multi-domain sentiment analysis, addressing data scarcity by leveraging cross-domain datasets. The Bidirectional Independent Recurrent Neural Network (Bi-IndRNN) Capsule model combined Bi-GRU for feature extraction with CapsuleNet to enhance performance. The model allowed each recurrent neuron to operate independently and achieved an average accuracy of 0.9144 on the Digikala dataset. Research by [14] optimized Indonesian text sentiment analysis by integrating feature selection (FS) using a genetic algorithm (GA) with deep learning models. The GA reduced data dimensionality from 41,140 to 20,769 features, improving accuracy and reducing computation time by 50%. Among the deep learning models, LSTM achieved the highest accuracy of 91.41%.

A work presented in [15] proposed a hybrid sentiment analysis model named DistilRoBiLSTMFuse, designed to extract deep contextual information from complex sentences for accurate sentiment identification. The model was evaluated on two benchmark datasets, IMDb and Twitter USAirline Sentiment, using preprocessing techniques such as data cleaning, customized stopword lists, and lemmatization. Oversampling was also employed to mitigate class imbalance. The model achieved 98.91% accuracy on IMDb and 99.42% on the Twitter dataset, surpassing previous models. The study [16] explored the application of deep learning techniques, particularly Long Short-Term Memory (LSTM), for sentiment analysis of user reviews from the Digikala website and Snapfood application. FastText-BiLSTM-CNN and Word2Vec-BiLSTM-CNN models were compared. The FastText-BiLSTM-CNN model achieved accuracies of 0.9754 and 0.8765 on datasets D1 and D2, respectively, outperforming the individual BiLSTM and CNN models.

According to [17], a dataset of 60,000 informal and conversational Persian texts was constructed, and a CNN-based sentiment analysis model was introduced. The model achieved 72% accuracy, representing a significant improvement in sentiment classification of Persian conversational texts. A separate investigation [18] analyzed Twitter sentiment using various machine learning models, including Vader, XGBoost, Random Forest, and LSTM variants. The goal was to classify tweets as positive, negative, or neutral to understand public opinion on global events. The Bidirectional LSTM model achieved the highest accuracy of 73%.

A recent study [19] introduced a hybrid framework for Persian sentiment analysis that integrated dependency-grammar–based linguistic rules with deep neural networks to overcome the limitations of traditional co-occurrence-based methods. The framework operated in a dual-mode configuration: when syntactic patterns were detected, symbolic dependency relations were leveraged to propagate sentiment from words to higher-level concepts;

otherwise, a sub-symbolic deep learning classifier was used. Experimental results on benchmark Persian product and hotel review datasets showed that this hybrid model outperformed traditional classifiers (e.g., SVM, Logistic Regression) by 10–15% and deep learning models (e.g., LSTM, CNN) by 3–4%.

Another contribution [20] introduced Senti-Persian, a large-scale, human-annotated sentiment dataset specifically designed for Persian, comprising 67,743 user-generated comments from Iranian streaming platforms (Namava, Filimo, Aparat) and social media (YouTube, Twitter, Instagram). To address data scarcity, a GAN-based synthetic data augmentation method tailored to Persian linguistic features was implemented. The augmented dataset improved classification accuracy from 88.4% to 96%.

Research presented in [21] investigated the performance of machine learning (ML) and deep learning (DL) models for Aspect-Based Sentiment Analysis (ABSA) on product reviews, utilizing Latent Dirichlet Allocation (LDA) for topic modeling. Using the Amazon review dataset, models such as Naïve Bayes, SVM, Random Forest, LSTM, and GRU were evaluated. Random Forest achieved the highest accuracy (94.5%) and F1-score (95.45%) for the reliability aspect, emphasizing the need to balance predictive accuracy with computational efficiency. The paper [22] proposed Depress-HybridNet, a hybrid deep learning framework for early depression detection on social media, integrating linguistic embeddings (BERT-BiLSTM) with behavioral activity patterns. Using an adaptive attention-based fusion mechanism, the model achieved an F1-score of 0.92 and an AUC of 0.93 on the Kaggle Depression Dataset. Clinical validation confirmed strong agreement with expert evaluations, highlighting the potential of hybrid modeling for predicting psychological states.

Another study [23] analyzed public sentiment in YouTube comments on the 2024 Indonesian presidential debate. The method combined BERT-based labeling, random oversampling, and a Multinomial Naïve Bayes classifier to address data imbalance and linguistic variability. The model achieved 85.15% accuracy, with an AUC of 96.8% and rapid classification time (0.000998 seconds), demonstrating effectiveness for real-time sentiment monitoring in political contexts. Finally, a hybrid deep learning model proposed in [24] combined BiGRU and Bi-Directional Attention Flow (BiDAF) for aspect-based sentiment analysis in e-commerce reviews. BiGRU captured sequential dependencies, while BiDAF enhanced attention on sentiment-relevant text segments. Trained on a preprocessed Amazon review dataset, the model achieved a peak training accuracy of 99.78%. Evaluation across four aspects—price, quality, service, and delivery—yielded F1-scores ranging from 0.90 to 0.92. Testing on the SemEval 2014 dataset demonstrated strong generalizability, achieving F1-scores of 88.78% for the restaurant domain and 83.66% for the laptop domain.

After reviewing these major contributions, it is evident that a variety of machine learning and deep learning techniques have been employed for sentiment analysis in Persian and other languages. These approaches differ in model architecture, dataset size, and domain coverage, leading to varying performance outcomes. As summarized in Table 1, the reported accuracies range from 72% to 99.7%, depending on the dataset and methodological complexity. The table provides a comparative overview of recent sentiment analysis models, highlighting their datasets, techniques, and key results.

Table 1. Comparison of sentiment analysis models and their performance across different datasets

Ref	Method	Dataset	Results
[9]	Stacked bidirectional LSTM / 2D-CNN	Movie reviews / Hotel reviews	Accuracy: 95.61%; 89.76%
[10]	CNN + LSTM	Persian multimodal dataset	Accuracy: 91.39%
[11]	CNN + fastText	*Digikala* online store	AUC: 0.996
[12]	Deep Learning (BERT)	Persian tweets	Accuracy: 97.7%
[13]	Bi-IndRNN Capsule Model	*Digikala* data	Accuracy: 91.44%
[14]	LSTM + Genetic Algorithm	Indonesian text	Accuracy: 91.41%
[15]	DistilRoBiLSTMFuse	IMDb / Twitter USAirline	Accuracy: 98.91%; 99.42%
[16]	FastText-BiLSTM-CNN	*Digikala* / Snapfood reviews	Accuracy: 97.54%; 87.65%
[17]	CNN-based model	60,000 informal Persian texts	Accuracy: 72%
[25]	Bidirectional LSTM	Twitter data	Accuracy: 73%

As observed in Table 1, while deep learning models tend to achieve high accuracy, they often rely on extensive labeled datasets and computationally demanding architectures. This motivates the adoption of hybrid feature-based approaches, such as the method proposed in this study, to achieve comparable accuracy with lower computational cost.

3. Proposed Method

The primary goal of this implementation is to address the challenge of sentiment analysis in Persian texts. Specifically, the system is designed to develop a machine learning model capable of classifying user reviews into two categories—positive and negative—based solely on the textual content of the reviews. This task represents a significant problem in the field of Natural Language Processing (NLP) with wide-ranging applications in customer feedback analysis, social media monitoring, user satisfaction management, and other opinion-mining systems.

In this framework, the data are first loaded from a text file, and after initial preprocessing, user reviews are categorized into positive (scores above 7) and negative (scores below 6) cases. Reviews with neutral or mid-range scores are excluded to define a clear binary classification problem. Subsequently, standard preprocessing techniques—such as removing punctuation marks, digits, and Persian-specific stopwords—are applied to prepare the textual data for analysis. Feature extraction is then performed using the TF-IDF method to vectorize the texts. Multiple machine learning models are trained and evaluated, and their performances are assessed using metrics such as Accuracy, Precision, Recall, and F1-score to determine the most effective configuration for Persian sentiment analysis.

Building on this foundation, the proposed method introduces an innovative framework that simultaneously leverages linguistic and structural information in Persian texts. This approach consists of two parallel feature extraction paths: one based on TF-IDF vectors that capture semantic meaning and word usage, and another that uses a set of handcrafted statistical features that describe writing style and lexical structure. The fusion of these feature domains is integrated within an SVM–based architecture, augmented with automated hyperparameter tuning. This dual-path design enables the model to combine surface-level stylistic patterns with deeper semantic signals, improving robustness and generalization. Figure 1 illustrates the overall workflow of the proposed system, highlighting the sequential stages from raw text preprocessing to feature extraction, feature fusion, and final classification.

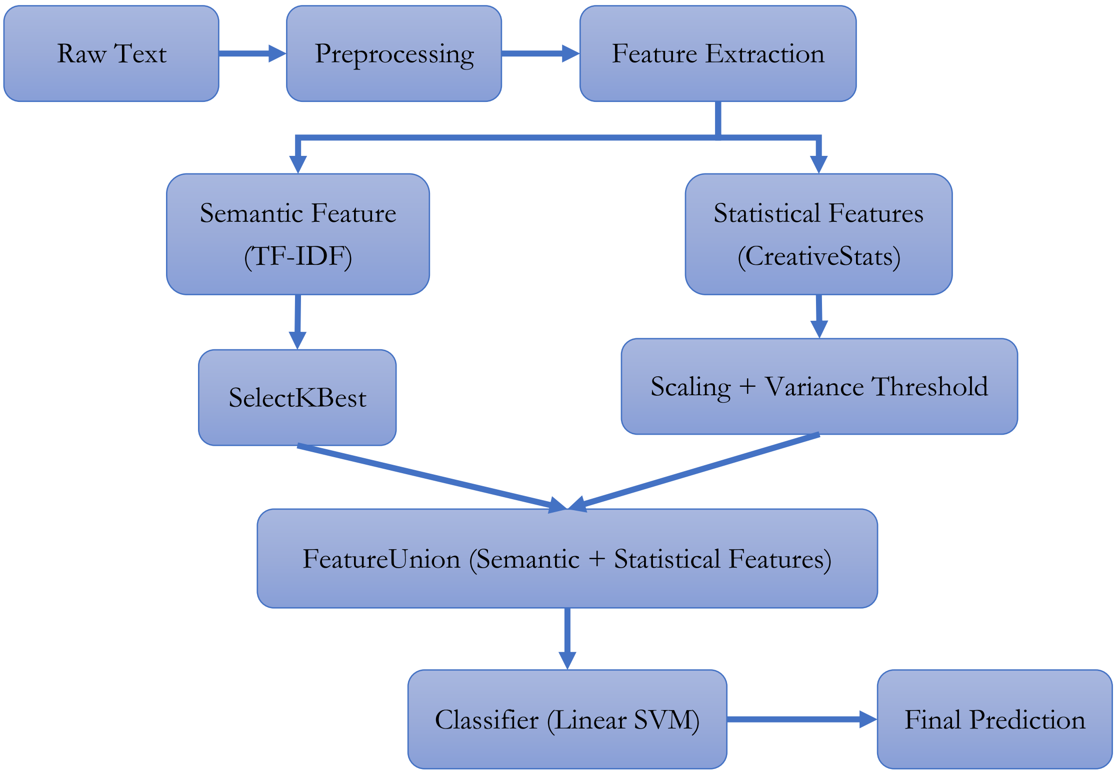

Figure 1. Flowchart of proposed system (TF-IDF + Statistical Features)

3.1. Text Preprocessing: Preparing Persian Text Data

Given that Persian texts—especially user-generated content—often contain non-standard punctuation, elongated or repetitive words, and irrelevant tokens, the preprocessing phase is crucial to ensure data consistency and clarity. At this stage, the data are cleaned to remove noise that could distort sentiment patterns.

As shown in Figure 2, the preprocessing process includes three main steps:

- Removal of all punctuation marks, numbers, and Persian-specific symbols such as "؟".
- Application of a hybrid stopword list that combines common and custom stopwords, effectively filtering uninformative words while retaining emotionally or descriptively meaningful ones.
- Filtering short words shorter than two characters to reduce sparsity and eliminate isolated tokens that do not convey sentiment.
- Additionally, character normalization and spacing are applied to handle informal variations in Persian text (e.g., different forms of "مي" or "می"). This ensures that equivalent tokens are treated uniformly across the corpus, improving feature reliability.

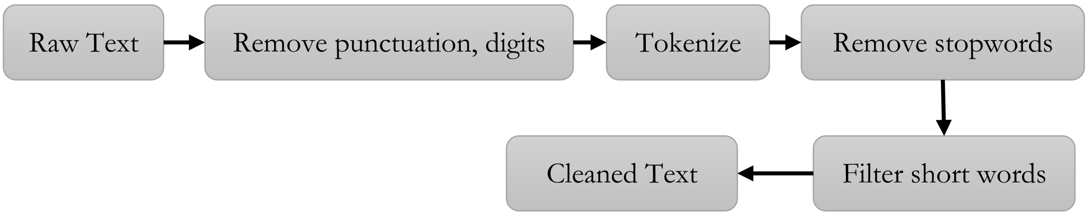

Figure 2. Text Preprocessing Steps

3.2. Semantic Representation via TF-IDF Vectors

In the first feature extraction path, each text is represented as a TF-IDF (Term Frequency–Inverse Document Frequency) vector, which quantifies the importance of words within a corpus. To capture contextual richness, n-grams ranging from unigrams to trigrams (1–3) are employed, enabling the detection of idiomatic expressions and multi-word sentiment phrases. A chi-squared statistical test is applied to select the top 3,000 most discriminative features, reducing dimensionality and computational overhead. This feature selection ensures that only the most relevant terms contributing to sentiment polarity are retained. Moreover, TF-IDF provides a sparse and interpretable representation, making it ideal for SVM-based models. Unlike embeddings that require extensive corpora, TF-IDF remains robust in low-resource scenarios such as Persian. This vectorization step therefore, serves as the backbone of semantic encoding for the proposed framework.

3.3. Statistical and Stylistic Text Features

Complementing the semantic path, the second extraction path focuses on statistical and stylistic cues that reflect surface-level linguistic behaviors. These features, generated by the CreativeStats module, include metrics that quantify an author's writing style, intensity, and expressive tone.

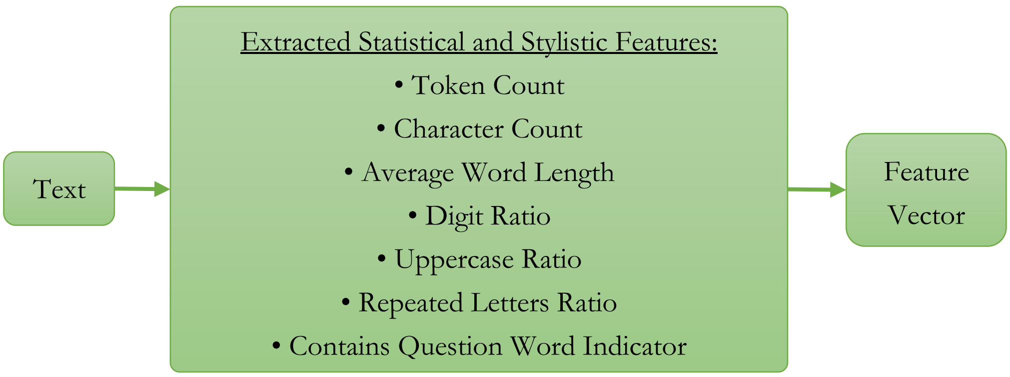

Figure 3. Creative statistics diagram illustrating the handcrafted statistical and stylistic features extracted from text.

As presented in Figure 3, the key features extracted are:

- Word and character counts: Indicators of text verbosity and engagement.
- Average word length: Reflecting complexity; emotional texts often use simpler words.
- Ratio of numeric characters: Useful for detecting enumerations, codes, or scores.
- Ratio of uppercase letters: While rare in Persian, may indicate copied or emphasized text.
- Character repetition ratio: Detects exaggerated emphasis common in informal Persian.
- Presence of interrogatives: e.g., "چرا" (why), "آیا" (whether), and "چطور" (how), often signaling critical sentiment.

To ensure consistency, these features are standardized using StandardScaler and filtered with a variance threshold to remove low-informative components. This step aligns the scale of the statistical features with the TF-IDF values, preventing any particular domain from dominating during feature fusion.

3.4. Feature Fusion: Combining Semantic and Statistical Features

Using FeatureUnion, the outputs of the two parallel feature extraction paths are concatenated into a single unified feature vector. This integration allows the model to simultaneously exploit the semantic content of text (via TF-IDF) and structural-linguistic information (via statistical and stylistic features), thereby creating a complementary and enriched representation that enhances performance in real-world applications. The overall integration process is illustrated in Figure 1, which shows how both feature branches are combined before classification.

In the proposed pipeline, two distinct feature representations are fused: (1) TF-IDF features extracted from text, and (2) handcrafted statistical features computed from sentence-level statistics using CreativeStats. The FeatureUnion module from scikit-learn concatenates the feature vectors from these two branches, forming a single combined feature vector for each text instance. This unified representation strengthens the model's ability to capture both semantic and stylistic cues that contribute to sentiment polarity.

To balance the contribution of each feature domain, explicit weighting factors α and β are introduced for the TF-IDF and statistical features, respectively. Mathematically, if the TF-IDF feature vector for an instance x is denoted by $f_1(x) \in \mathbb{R}^m$ and the handcrafted feature vector by $f_2(x) \in \mathbb{R}^n$, the fused feature vector $f(x)$ is defined as Equation (1).

$$f(x) = [\alpha \cdot f_1(x) \| \beta \cdot f_2(x)] \tag{1}$$

Here, α and β control the relative influence of each feature domain. The resulting weighted feature vector $f(x)$ is then fed into the LinearSVC classifier for training and prediction. The weighting parameters were optimized automatically using RandomizedSearchCV, along with the SVM hyperparameter C, to achieve the best trade-off between semantic and statistical contributions.

Since TF-IDF vectors are inherently L2-normalized by scikit-learn, no additional scaling is required for the semantic feature branch. However, the handcrafted statistical features operate on different numerical ranges. To ensure comparability and prevent dominance of large-magnitude statistical values in the optimization process, these features are normalized using StandardScaler before fusion. This step stabilizes the hybrid feature space and ensures balanced contribution from both semantic and stylistic components. Overall, the fusion mechanism effectively leverages complementary information from both feature types. By integrating semantic signals from TF-IDF with handcrafted cues such as repetition, word length, and numeric ratios, the model can better capture emotional tone and stylistic emphasis in informal Persian text. As shown in Figure 1, this unified representation serves as the foundation of the proposed architecture, leading to improved sentiment classification accuracy across both positive and negative classes.

3.5. Classification and Hyperparameter Optimization

For final classification, a Linear Support Vector Classifier (LinearSVC) is employed due to its effectiveness in handling high-dimensional, sparse data typical of TF-IDF representations. The model's hyperparameters, including the regularization parameter C, are optimized using RandomizedSearchCV with 5-fold cross-validation to achieve an optimal trade-off between margin maximization and generalization performance.

A linear kernel is specifically chosen to reduce computational overhead while maintaining high accuracy in separable feature spaces. Prior research [17] has demonstrated that linear SVMs perform competitively on text classification tasks where semantic features exhibit near-linear separability. Additionally, StandardScaler normalization is retained to ensure comparability between semantic and statistical features, preventing numerical imbalance that could bias the classifier. This systematic optimization yields a stable, interpretable, and computationally efficient sentiment classifier suited for medium-sized Persian text datasets.

3.6. Method Summary and Workflow Clarification

To provide a clearer picture of the proposed methodology, the overall workflow integrates preprocessing, dual-path feature extraction, feature fusion, and classification into a unified and systematic pipeline. Each stage in this framework has been carefully designed to address specific challenges in Persian sentiment analysis. The preprocessing stage removes noise and irrelevant tokens to ensure meaningful input data. The semantic extraction path, based on TF-IDF, captures the contextual and frequency-based significance of words that reflect sentiment polarity. In parallel, the statistical and stylistic path models the author's writing behavior, punctuation usage, and linguistic variations that often carry implicit emotional cues.

As shown in Figure 1, both paths produce feature vectors that are fused using FeatureUnion into a single, enriched representation, which is then fed into the Linear SVM classifier optimized through RandomizedSearchCV. This integration allows the system to exploit both semantic and surface-level features effectively, improving classification robustness and interpretability. To evaluate the effectiveness of the proposed framework, several standard performance metrics were used, namely:

- Accuracy, which measures the overall proportion of correctly classified samples and provides a general view of the model's predictive power.
- Precision, which indicates how many of the samples identified as positive are indeed positive, reflecting the model's reliability in prediction.
- Recall, which measures the proportion of actual positive samples correctly identified by the classifier, showing the model's sensitivity to positive sentiment.
- F1-Score, which represents the harmonic mean of Precision and Recall, providing a balanced indicator that remains informative even when class distributions are uneven.

These metrics are widely adopted in sentiment analysis research because they collectively describe the balance between precision and sensitivity, enabling an objective comparison with baseline models. The reported results are therefore expressed in terms of these four key evaluation measures to ensure transparency and reproducibility of the findings. Furthermore, the choice of Linear SVM with automatic hyperparameter tuning contributes to the model's computational efficiency and stability, allowing consistent performance across different configurations of the feature space.

4. Experiments and Results

To validate the effectiveness of the proposed hybrid sentiment analysis framework, a series of controlled experiments were conducted. This section describes the experimental environment, dataset characteristics, and model configuration used during the evaluation. The goal of these experiments is to assess the performance of the proposed feature fusion approach in comparison with conventional machine learning classifiers under consistent and reproducible conditions.

4.1. Experimental Environment

All experiments were implemented in Python 3.10, using the scikit-learn 1.3, NumPy 1.24, pandas 2.0, and SciPy 1.10 libraries. The system was executed on a workstation equipped with an Intel Core i7 processor, 16 GB of RAM, and Windows 10 (64-bit) operating system. Since the proposed method relies on classical machine learning algorithms rather than deep neural architectures, GPU resources were not required. The entire workflow—including preprocessing, feature extraction, feature fusion, and classification—was executed within the scikit-learn pipeline environment to ensure full reproducibility. The use of a deterministic random seed across all experiments also guarantees that results can be replicated under identical conditions. This experimental setup provides a stable and transparent computational environment suitable for validating the performance of the proposed SVM-based framework.

4.2. Dataset Description

The dataset used in this study was collected from Digikala.com, one of the largest Persian e-commerce platforms containing user-generated product reviews. Each record includes three attributes:

- Text: the content of the user's comment, written in Persian.
- Score: a numeric rating from 0 to 100 representing user satisfaction.
- Suggestion: an optional indicator of purchase recommendation (1 = recommend, 2 = not recommend, 3 = neutral).

To generate sentiment labels, the Score column was used according to the following rules:

- Positive sentiment: Score > 70
- Negative sentiment: Score < 60
- Neutral or discarded samples: 60 ≤ Score ≤ 70

All neutral comments were excluded to maintain a clear binary classification task. After filtering, the dataset contained 1,687 positive and 499 negative samples, yielding a total of 2,186 labeled records. Each comment was preprocessed following the pipeline described in Section 3.1. The dataset represents a realistic domain of informal product reviews, where text variability and noise make classification challenging. This setting effectively tests the robustness of combining textual and statistical features. To ensure fair evaluation, the dataset was divided into training and testing subsets using a stratified train–test split that preserves class distribution. A fixed random seed (random_state = 42) was applied to maintain reproducibility. Table 2 summarizes the dataset's key statistics, including total data points, average text length, and sentiment distribution.

Table 2. Dataset statistics.

Number of Data Points	Avg. Text Length (Characters)	Avg. Text Length (Words)	Positive Comments	Negative Comments
3,200	247.96	51.60	1,687	499

This dataset provides a solid foundation for analyzing Persian sentiment due to its real-world origin, moderate imbalance, and diverse vocabulary, which collectively allow for rigorous evaluation of hybrid feature integration techniques.

4.3. Model Configuration and Training Setup

The overall system integrates all preprocessing, feature extraction, and classification components into a unified scikit-learn pipeline, ensuring modularity and consistency. The configuration of the proposed model is summarized in Table 3, which details the preprocessing operations, feature extraction modules, architectural components, training configuration, and classifier settings.

The pipeline consists of two parallel feature branches—TF-IDF and handcrafted statistical features—that are combined using FeatureUnion and passed to a Linear SVM classifier. Hyperparameter optimization is performed using RandomizedSearchCV to automatically identify the optimal value for the regularization parameter (C). A linear kernel was selected for its computational efficiency and strong performance in high-dimensional, sparse feature spaces. Both TF-IDF and handcrafted features were normalized and scaled to ensure compatibility within the same vector space.

This configuration ensures reproducibility and interpretability by explicitly detailing all hyperparameters and preprocessing operations. The transparent documentation of both feature branches and training setup allows other researchers to replicate or extend the proposed system for future work in Persian sentiment analysis.

Table 3. Model configuration.

Component	Description	Parameters / Settings
Preprocessing	Custom text cleaning and token filtering	• Remove punctuation and digits (Python translate) • Remove Persian stopwords (stopwords-fa.txt, spaCy, and custom list) • Token length > 1 • Lowercase preserved (no case normalization)
Feature Extraction	Combination of TF-IDF text features and handcrafted statistical features	TF-IDF: • max_features = 12,000 • ngram_range = (1,3) • SelectKBest(χ^2, k = 3,000) Handcrafted Features: • Number of words • Number of characters • Average word length • Digit ratio • Uppercase ratio • Repeated-character ratio • Question-word indicator • Standardization (StandardScaler) • Variance threshold = 0.01
Model Architecture	Unified scikit-learn pipeline combining preprocessing, feature union, and classifier	1. Preprocess() 2. FeatureUnion([TF-IDF branch, CreativeStats branch]) 3. LinearSVC()
Training Configuration	Randomized hyperparameter search over SVM parameter C	RandomizedSearchCV with: • param_dist = {'clf__C': uniform(0.01, 10)} • n_iter = 20 • cv = 5 • scoring = 'accuracy' • random_state = 42
Classifier	Linear Support Vector Machine (SVM)	• LinearSVC() • Optimal C = search.best_params_['clf__C']

4.4. Results and Discussion

To evaluate the proposed hybrid sentiment analysis system, its performance was compared against four baseline classifiers: Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree. All baseline models used only TF-IDF features, while the proposed model employed both TF-IDF and statistical-stylistic features, integrated within an optimized linear SVM framework using RandomizedSearchCV. Evaluation metrics included precision, recall, F1-score, and accuracy, as shown in Tables 4–6.

4.4.1. Performance on Negative and Positive Sentiment Classes

The experimental results demonstrate notable differences in model behavior between positive and negative sentiment detection. As shown in Table 4 and Figure 4, most traditional models performed well in identifying positive reviews, but struggled to accurately detect negative sentiment, a common challenge in imbalanced Persian text datasets.

Naïve Bayes and Logistic Regression models, though achieving perfect recall (1.00) for positive samples, completely failed to identify negative samples (F1 = 0.0). This indicates severe bias toward the majority class, resulting in high false positives. In contrast, Decision Tree and Random Forest models exhibited more balanced performance, but their recall for negative samples remained below 0.50, showing limited sensitivity to minority-class expressions.

As illustrated in Table 5 and Figure 5, all models performed considerably better in recognizing positive sentiments, reflecting the class imbalance within the dataset. The proposed system, by fusing TF-IDF and handcrafted statistical features, achieved the best overall balance. It reached an F1-score of 0.53 for negative reviews and 0.90 for positive reviews, indicating a significant improvement in identifying underrepresented sentiment patterns. The combination of semantic and structural cues allowed the model to better interpret stylistic indicators, such as punctuation and repetition, that are often associated with negative emotions in Persian informal text.

Table 4. Performance comparison of models in negative sentiment detection.

Model	F1-score	Recall	Precision
Naïve Bayes + TF-IDF	0.00	0.00	0.00
Logistic Regression + TF-IDF	0.06	0.03	1.00
Random Forest + TF-IDF	0.34	0.21	0.90
Decision Tree + TF-IDF	0.49	0.46	0.51
Proposed System	0.53	0.40	0.77

Table 5. Performance comparison of models in positive sentiment detection.

Model	F1-score	Recall	Precision
Naïve Bayes + TF-IDF	0.87	1.00	0.77
Logistic Regression + TF-IDF	0.87	1.00	0.78
Random Forest + TF-IDF	0.89	0.99	0.81
Decision Tree + TF-IDF	0.86	0.87	0.85
Proposed System	0.90	0.96	0.84

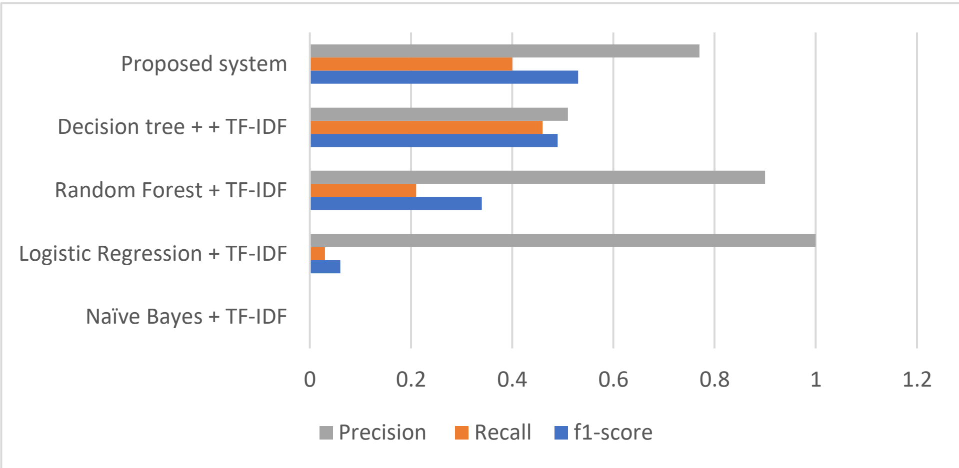

Figure 4. Model performance comparison for negative class classification.

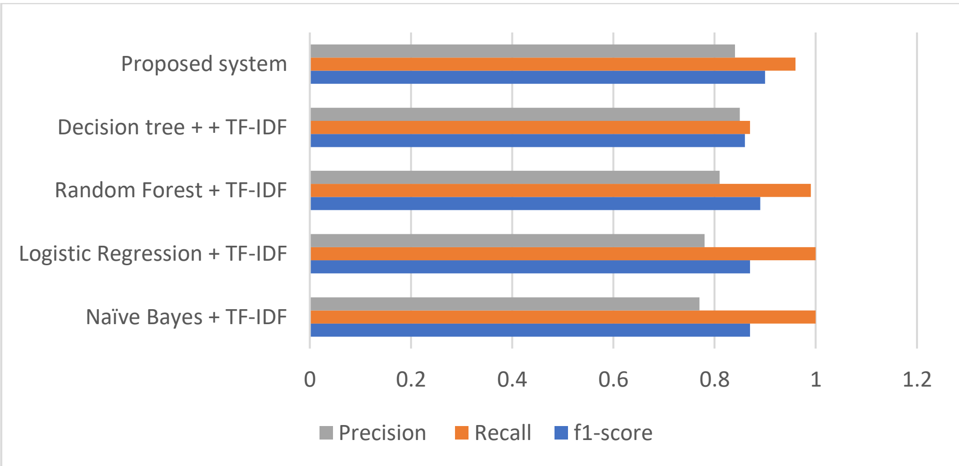

Figure 5. Model performance comparison for positive class classification.

4.4.2. Overall Accuracy Comparison

The aggregated results in Table 6 and Figure 6 confirm the superior overall performance of the proposed hybrid model. It achieved the highest accuracy of 0.8354, outperforming all baseline models—Random Forest (0.8135), Decision Tree (0.777), Logistic Regression (0.7788), and Naïve Bayes (0.7715). This improvement highlights the effectiveness of incorporating handcrafted statistical and stylistic features alongside conventional TF-IDF text representations. Such hybridization enables the model to capture subtle expressive signals, particularly in informal Persian text, where emotional emphasis is often conveyed through orthographic cues rather than explicit vocabulary.

Table 6. Overall model accuracy comparison.

Model	Accuracy
Naïve Bayes + TF-IDF	0.7715
Logistic Regression + TF-IDF	0.7788
Decision Tree + TF-IDF	0.7770
Random Forest + TF-IDF	0.8135
Proposed System	0.8354

4.4.3. Discussion

The proposed model demonstrates consistent advantages over conventional TF-IDF–based classifiers. By integrating linguistic, statistical, and structural features, it effectively mitigates bias toward dominant classes and enhances the detection of minority (negative) sentiments. The performance gain of approximately 2.2% in accuracy over the best baseline (Random Forest) confirms the practical value of feature fusion for Persian sentiment analysis. Moreover, the use of RandomizedSearchCV for hyperparameter tuning yielded stable, generalizable results, indicating that the performance improvement was not due to overfitting.

Although deep learning models may achieve higher accuracy on very large corpora, the proposed system offers a computationally efficient, interpretable, and data-efficient alternative for medium-scale Persian datasets. Overall, these findings validate the robustness of hybrid feature engineering in low-resource languages, where data scarcity and stylistic variation pose unique challenges for sentiment analysis.

4.5. Ablation Study

To further examine the contribution of each feature domain, an ablation study was conducted focusing on three configurations of the Support Vector Machine (SVM) classifier: (a) SVM + TF-IDF only, (b) SVM + statistical features only, and (c) SVM + combined features (the proposed model). This analysis was restricted to SVM because it serves as the core classifier in the proposed system and demonstrated the most stable and interpretable results in earlier experiments. Additionally, SVM's linear separability and robustness to high-dimensional sparse features make it an appropriate choice for analyzing the relative effects of feature fusion. The results, presented in Table 7, highlight the performance variations across these three configurations.

Table 7. Performance of SVM using different feature components.

Configuration	Accuracy	F1 (Negative)	F1 (Positive)	Precision (Negative)	Precision (Positive)	Recall (Negative)	Recall (Positive)
SVM + TF-IDF only	0.78	0.48	0.88	0.50	0.87	0.46	0.90
SVM + Statistical only	0.80	0.42	0.85	0.45	0.84	0.40	0.87
SVM + Combined (Proposed)	0.8354	0.53	0.90	0.77	0.84	0.40	0.96

As shown in Table 7, using TF-IDF alone yields reasonable performance for the positive class (F1 = 0.88) but moderate performance for the negative class (F1 = 0.48). Similarly, statistical features alone yield acceptable results for positive samples (F1 = 0.85), but remain limited for negative reviews (F1 = 0.42), confirming that stylistic indicators alone cannot fully capture sentiment polarity.

In contrast, the combined feature configuration achieves the highest overall accuracy (0.8354) and consistent improvements across nearly all metrics. Notably, the precision for negative reviews increases markedly to 0.77, while the model maintains robust performance for positive reviews (F1 = 0.90, recall = 0.96). These findings indicate that integrating statistical and textual domains enhances both balance and discriminative power, particularly for minority classes.

The improvement can be attributed to the complementary nature of the features: statistical attributes such as repeated-character ratio, digit ratio, uppercase usage, average word length, and question-word presence help capture emotional and stylistic nuances that are often missed by TF-IDF representations. For instance, repeated letters may signify emphasis or sarcasm, numerical tokens can express evaluative intensity, and interrogative forms may signal uncertainty or dissatisfaction.

By incorporating these cues, the hybrid SVM model becomes more sensitive to the informal, expressive nature of Persian user reviews, enabling it to better recognize both positive and negative sentiments. Overall, this ablation analysis confirms that integrating semantic and stylistic information is essential for achieving reliable, interpretable sentiment classification in low-resource Persian text datasets.

5. Conclusions

In this study, a novel approach for sentiment analysis of Persian text was proposed by combining textual (TF-IDF) and handcrafted statistical features with an optimized Support Vector Machine (SVM) classifier. The main objective of this research—to develop an effective Persian sentiment analysis model capable of handling imbalanced data—was successfully achieved. The proposed system was evaluated against four baseline models (Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree), all relying solely on TF-IDF features. Experimental results demonstrate that the proposed system outperforms these baselines across multiple metrics, including overall accuracy, F1-score, recall, and precision. Specifically, the model achieved the highest accuracy (0.8354) and substantially improved performance in identifying negative reviews, addressing the class imbalance issue observed in baseline models.

The ablation study further confirmed the complementary nature of textual and statistical features. Statistical indicators such as repeated characters, digit ratios, uppercase usage, average word length, and question-word indicators capture informal emotional patterns that TF-IDF alone may overlook. Their integration enhances the model's ability to detect subtle sentiment cues, particularly in informal Persian product reviews, and explains the observed improvements in F1 and precision for the negative class. While effective for informal review data, the proposed method may have limited applicability for formal texts such as news articles or academic writing, where informal stylistic cues are less frequent. Future research could address this limitation by redesigning or extending the feature set to improve generalization across domains.

Although this approach was developed for Persian, its framework—combining TF-IDF representations with handcrafted statistical features—can be adapted to other morphologically rich languages with minimal modifications. Such adaptation would primarily involve language-specific preprocessing (e.g., tokenization, stopword lists) and tailoring handcrafted features to reflect each language's unique structural or stylistic patterns. Direct comparison with deep learning models is limited due to differences in datasets and labeling schemes. However, the proposed TF-IDF + statistical feature approach demonstrates strong and consistent performance (accuracy = 0.8354) on the Digikala dataset while remaining computationally efficient and interpretable. This makes it practical for resource-constrained or small-scale text analysis scenarios.

In future work, expanding the dataset to include additional Persian sources, exploring advanced feature engineering or ensemble techniques, and integrating modern embeddings could further improve model performance. Overall, the results demonstrate that combining handcrafted statistical features with textual features and optimized machine learning classifiers can significantly enhance sentiment analysis, particularly in imbalanced and informal text datasets.

Funding: This research received no external funding.

Data Availability Statement: The dataset utilized in this study was derived from user comments on the Digikala e-commerce website and processed specifically for research purposes. Data supporting the findings of this study are available from the corresponding author upon reasonable request. The dataset can also be accessed at https://github.com/research2025-arch/sentiment_analysis_dataset.

Acknowledgments: The author would like to thank all individuals who provided support and guidance during the preparation of this research.

Conflicts of Interest: The author declares no conflict of interest.

References

[1] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, no. 7, p. 4550, Apr. 2023, doi: 10.3390/app13074550.

[2] R. Das and T. D. Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–38, Dec. 2023, doi: 10.1145/3586075.

[3] K. K. Yusuf, E. Ogbuju, T. Abiodun, and F. Oladipo, "A Technical Review of the State-of-the-Art Methods in Aspect-Based Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 287–298, Feb. 2024, doi: 10.62411/jcta.9999.

[4] Z. Rajabi and M. Valavi, "A Survey on Sentiment Analysis in Persian: a Comprehensive System Perspective Covering Challenges and Advances in Resources and Methods," *Cognit. Comput.*, vol. 13, no. 4, pp. 882–902, Jul. 2021, doi: 10.1007/s12559-021-09886-x.

[5] F. Amiri, S. Scerri, and M. H. Khodashahi, "Lexicon-based Sentiment Analysis for Persian Text," in *Proceedings of Recent Advances in Natural Language Processing*, 2015, pp. 9–16. [Online]. Available: https://aclanthology.org/R15-1002.pdf

[6] M. Shamsfard, "Towards Semi Automatic Construction of a Lexical Ontology for Persian," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008, pp. 2629–2633. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/654_paper.pdf

[7] M. E. Basiri, N. Ghasem-Aghaee, and A. R. Naghsh-Nilchi, "Lexicon-based Sentiment Analysis in Persian," in *Current and Future Developments in Artificial Intelligence*, BENTHAM SCIENCE PUBLISHERS, 2017, pp. 154–183. doi: 10.2174/9781681085029117010009.

[8] P. Monika, C. Kulkarni, N. Harish Kumar, S. Shruthi, and V. Vani, "Machine learning approaches for sentiment analysis," *Int. J. Health Sci. (Qassim).*, pp. 1286–1300, Apr. 2022, doi: 10.53730/ijhs.v6nS4.6119.

[9] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment Analysis of Persian Movie Reviews Using Deep Learning," *Entropy*, vol. 23, no. 5, p. 596, May 2021, doi: 10.3390/e23050596.

[10] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for persian sentiment analysis," *Neurocomputing*, vol. 457, pp. 377–388, Oct. 2021, doi: 10.1016/j.neucom.2021.02.020.

[11] S. Shumaly, M. Yazdinejad, and Y. Guo, "Persian sentiment analysis of an online store independent of preprocessing using convolutional neural network with fastText embeddings," *PeerJ Comput. Sci.*, vol. 7, p. e422, Mar. 2021, doi: 10.7717/peerj-cs.422.

[12] M. Dehghani, D. T. Dehkordy, and M. Bahrani, "Abusive words Detection in Persian tweets using machine learning and deep learning techniques," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, Dec. 2021, pp. 1–5. doi: 10.1109/ICSPIS54653.2021.9729390.

[13] R. Mousa, M. A. Dadgostarnia, A. O. Malamiri, E. Behnam, and A. Mohammadi, "The IndRNN Capsule Approach for Persian Multi-Domain Sentiment Analysis," *J. Appl. Res. Ind. Eng.*, vol. 11, no. 3, pp. 436–454, 2024, doi: 10.22105/jarie.2023.345352.1477.

[14] S. Mujilahwati, N. Z. Mohd Safar, K. M. N. Ku Khalif, and N. Ghazalli, "Optimizing Sentiment Analysis of Indonesian Texts: Enhancing Deep Learning Models with Genetic Algorithm-Based Feature Selection," *J. Soft Comput. Data Min.*, vol. 5, no. 2, Dec. 2024, doi: 10.30880/jscdm.2024.05.02.016.

[15] S. K. Papia, M. A. Khan, T. Habib, M. Rahman, and M. N. Islam, "DistilRoBiLSTMFuse: an efficient hybrid deep learning approach for sentiment analysis," *PeerJ Comput. Sci.*, vol. 10, p. e2349, Sep. 2024, doi: 10.7717/peerj-cs.2349.

[16] F. Soleimanzadeh and M. J. Shayegan, "Sentiment Analysis in Persian Language Using Long Short-Term Memory and Deep Learning," in *2025 11th International Conference on Web Research (ICWR)*, Apr. 2025, pp. 597–603. doi: 10.1109/ICWR65219.2025.11006222.

[17] M. Mazoochi, L. Rabiei, F. Rahmani, and Z. Rajabi, "Constructing colloquial dataset for Persian sentiment analysis of social microblogs," *Multimed. Tools Appl.*, vol. 84, no. 33, pp. 41567–41582, Apr. 2025, doi: 10.1007/s11042-025-20777-3.

[18] M. Khan and A. Srivastava, "Sentiment Analysis of Twitter Data Using Machine Learning Techniques," *Int. J. Eng. Manag. Res.*, vol. 14, no. 1, pp. 2250–0758, 2024, doi: 10.5281/zenodo.10791485.

[19] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks," *Neurocomputing*, vol. 380, pp. 1–10, Mar. 2020, doi: 10.1016/j.neucom.2019.10.009.

[20] M. Mohammadi, M. R. Amin, and S. Tavakoli, "Boosting sentiment analysis in Persian through a GAN-based synthetic data augmentation method," in *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, 2025, pp. 54–63. [Online]. Available: https://aclanthology.org/2025.abjadnlp-1.7/

[21] D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, May 2024, doi: 10.62411/faith.2024-3.

[22] J. B. Oluwagbemi, A. E. Mesioye, and R. S. Akinbo, "Depress-HybridNet: A Linguistic-Behavioral Hybrid Framework for Early and Accurate Depression Detection on Social Media," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 432–444, Sep. 2025, doi: 10.62411/faith.3048-3719-266.

[23] A. Angdresey, L. Sitanayah, and I. L. H. Tangka, "Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 342–354, Jan. 2025, doi: 10.62411/jcta.11668.

[24] D. R. I. M. Setiadi, W. Warto, A. R. Muslikh, K. Nugroho, and A. N. Safriandono, "Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 470–480, Apr. 2025, doi: 10.62411/jcta.12376.

[25] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, 1998, pp. 137–142. doi: 10.1007/BFb0026683.