Business Intelligence Management Studies
Vol. 15, No.55, Spring 2026; P 55-84
ims.atu.ac.ir
DOI: 10.22054/ims.2026.87943.2666

ATU PRESS

Original Research
Accepted: 31 January 2026
Revised: 29 December 2025
Received: 07 October 2025
eISSN: 2821-0816
ISSN: 2821-0964

A Multilingual BERT Framework for Intelligent Screenplay Analysis: Emotion Recognition through Character Behavioral Patterns

Zohreh Jafarbeglou — PhD Candidate, Department of Information Technology Management, Qeshm Branch, Islamic Azad University, Qeshm, Iran

Mohammad Ali Afshar Kazemi * — Professor, Department of Industrial Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran

Soheila Jokar — Assistant Professor, Department of Mathematics and Statistics, Qeshm Branch, Islamic Azad University, Qeshm, Iran

Abstract

This study introduces EmoBERTScr, an intelligent Multilingual BERT-based model for detecting and analyzing emotional and behavioral patterns in Persian and multilingual screenplays. A dataset of 35,800 dialogue samples from 1,700 Persian and English screenplays was manually collected and annotated by linguistics and psychology experts into eight emotional categories according to Plutchik's (1980) wheel of emotions. The methodology involved data acquisition from cinematic archives, preprocessing (noise removal, normalization, and tokenization), and supervised learning with a fine-tuned Multilingual BERT architecture. The model achieved an overall accuracy of 98.22%, with F1-scores ranging from 95.53% (surprise) to 100% (joy, fear, sadness, trust) across all emotional categories. The primary contribution of this research lies in developing a practical AI assistant for screenwriters, capable of providing real-time feedback to enhance narrative coherence and maintain emotional consistency, which can reduce rewriting and production costs in the film industry. While certain overlapping emotional states, such as anticipation

This article is derived from the Ph.D. dissertation in Information Technology Management, Islamic Azad University, Qeshm Branch.

* Corresponding Author: moh.afsharkazemi@iauctb.ac.ir

How to Cite: Jafarbeglou, Z., Afshar Kazemi, M. Jokar. S. (2026). A Multilingual BERT Framework for Intelligent Screenplay Analysis: Emotion Recognition through Character Behavioral Patterns, *Journal of Business Intelligence Management Studies*, 15(56), 55-84. DOI: 10.22054/ims.2026.87943.2666

and surprise**,** remain challenging to distinguish, the proposed approach demonstrates robust performance in both dominant and subtle emotional expressions. This work establishes a foundation for future advancements in intelligent narrative analysis, particularly for low-resource languages like Persian, and highlights the potential for integrating AI-driven emotion recognition into professional screenplay development workflows.

Keywords: Multilingual BERT; EmoBERTScr; Emotion Recognition; Intelligent Assistant; Character Behavior; Persian Screenplays; Plutchik's Emotion Wheel; Narrative Coherence.

Introduction

Screenplays constitute the foundational core of the film industry, serving as the primary medium through which human emotions and behaviors are conveyed across diverse narrative genres such as drama, comedy, tragedy, and romance. Their pivotal role in audience engagement and commercial success has been extensively documented (Bordwell & Thompson, 2016). Analyzing emotional and behavioral patterns within these dialogue-driven texts enables data-driven approaches to film production and management by facilitating the assessment of narrative emotional impact, thereby supporting production optimization and informed strategic decision-making (Nandwani & Verma, 2021).

Persian cinema, deeply rooted in rich literary and cultural traditions, presents distinctive challenges for automated textual analysis. Linguistic characteristics such as complex morphology, implicit contextual meanings, culture-specific metaphors, and non-explicit emotional expressions pose significant obstacles for computational modeling. These challenges are further compounded by the scarcity of annotated datasets, which hinders the development of robust natural language processing solutions for Persian screenplays (Heydari et al., 2024).

Recent advances in natural language processing, particularly transformer-based architectures such as BERT with bidirectional contextual representation, have demonstrated strong capabilities in the analysis of creative texts (Rahmani Zardak et al., 2023). However, the application of such models to low-resource languages remains constrained by insufficient domain-specific training data. This limitation is especially pronounced in screenplay analysis, where conversational language, sarcasm, irony, rapid emotional shifts, and narrative dependencies differ substantially from the linguistic patterns found in commonly used training corpora (Dashtipour et al., 2021a).

A review of the literature reveals three critical research gaps that remain largely unaddressed. First, virtually no research has specifically targeted dialogue-rich cinematic screenplays in the Persian language. Second, existing sentiment analysis models have predominantly been trained on film reviews, social media content, or news articles, resulting in notable performance degradation when applied to screenplay-specific linguistic features. Third, no large-scale, publicly available annotated dataset of Persian screenplay dialogues has been released, thereby limiting reproducibility and benchmarking efforts in this domain.

To address these gaps, this study proposes a domain-adapted Multilingual BERT (mBERT) framework for the automated identification of emotional and behavioral patterns in Persian and multilingual screenplays. By leveraging manually annotated screenplay dialogues and fine-tuning mBERT for narrative-specific linguistic features, the study aims to enhance emotion recognition accuracy while providing analytical insights that support narrative coherence assessment and intelligent decision-making in film production and management contexts.

Research Questions

The primary research questions guiding this study are as follows:

1. How effectively can a fine-tuned Multilingual BERT (mBERT) model identify emotional and behavioral patterns in dialogue-rich Persian and multilingual screenplays?
2. To what extent do domain-specific, manually annotated screenplay datasets improve emotion recognition performance compared to existing Persian and multilingual baseline models?
3. How can automated emotion and behavior analysis of screenplays contribute to enhancing narrative coherence and supporting data-driven decision-making in intelligent film business management, particularly in low-resource language contexts such as Persian?

By manually collecting and annotating 35,800 dialogue samples from 1,700 Persian and English screenplays based on Plutchik's wheel of emotions, and by developing a domain-adapted mBERT architecture achieving an overall accuracy of 98.22%, this study provides a validated framework for automated emotion analysis in cinematic texts. The results offer empirical evidence for improving emotional coherence assessment and establish a data-driven foundation for future research and intelligent applications in narrative analysis and film industry decision support systems.

Literature Review

Screenplays, as dialogue-driven texts featuring rapid emotional shifts and culture-dependent contexts, pose unique challenges for sentiment analysis in natural language processing (Frangidis et al., 2020). These texts contain layered semantic structures arising from character interactions and linguistic nuances such as sarcasm and irony, which significantly complicate automated analysis (Bordwell & Thompson, 2016). Recent advances in NLP—particularly transformer-based models like BERT with bidirectional contextual processing—have improved the accuracy of cinematic text analysis and provided tools for data-driven management in the film industry (Nandwani & Verma, 2021; Devlin et al., 2019). For instance, Dashtipour et al. (2021a) reported F1-scores ranging from 85 to 90% for Persian short stories, although their focus on non-dialogue texts limited applicability to Persian screenplays. Similarly, Farahani et al. (2023) achieved 94% accuracy in Persian literary texts; however, the scarcity of dialogue-rich datasets constrained the utility of their models for screenplay analysis. Beyond dataset limitations, in low-resource languages such as Persian, morphological complexities (e.g., compound verbs) and the limited availability of annotated datasets further hinder automated analysis (Heydari et al., 2024). Dehghani et al. (2021) demonstrated the effectiveness of deep learning models on Persian literary texts but noted limitations when applied to screenplays due to their multilayered narrative structures. Dashtipour et al. (2021b) highlighted the importance of culture-aware models for analyzing Persian film reviews. Pólya and Csertő (2023), by examining

emotional narrative structures and achieving an F1-score of 0.72, showed that narrative architecture can effectively inform emotion recognition—a methodology potentially extendable to Persian screenplay analysis. Studies in Persian have also pointed to AI's potential in sentiment analysis and intelligent business management. For example, Khadivar et al. (2024) reported positive user attitudes toward technologies such as ChatGPT, while Mohebbi and Torfi (2025) and Valizadeh Hamzekolaei et al. (2024) emphasized the role of scenario planning and sentiment analysis in developing Iranian intelligent businesses. Automated analysis of screenplays enhances economic efficiency and narrative dynamics in Persian cinema by predicting emotional resonance (Dashtipour et al., 2021b; Heydari et al., 2024). Nevertheless, critical research gaps persist: most NLP research focuses on high-resource languages, there is insufficient analysis of dialogue-rich screenplays, and application of NLP in Persian cinema remains limited. To address these gaps, this study develops a Multilingual BERT model enhanced with culturally informed annotation, offering an intelligent tool for narrative enhancement and data-driven cinematic management.

Table 1. Summary of Literature Review

Author(s)	Research Title	Methodology	Key Findings
Ligthart et al. (2024)	Systematic reviews in sentiment analysis: a tertiary study	Systematic review of SA challenges and applications	Limitations in low-resource languages; need for annotated datasets
Heydari et al. (2024)	Deep Learning-based Sentiment Analysis in Persian	Design and implementation of a hybrid deep learning model with regularization techniques on Digikala product reviews	Achieved F1-score of 78.3 across three emotional classes
Pólya & Csertő (2023)	Emotion Recognition Based on the Structure of Narratives	Collection of emotional narratives; linguistic and structural analysis using rule-based algorithms and machine learning to predict emotional states of 117 participants	Hybrid approach achieved mean F1-score of 0.72; demonstrated benefits of narrative structure analysis in NLP-based emotion recognition
Farahani et al. (2023)	Persian Text Sentiment Analysis Based on BERT and Neural Networks	Sentiment analysis of Persian texts using BERT and neural networks	94% accuracy on Persian literary texts; limited applicability to screenplays due to scarcity of dialogue-rich data
Dashtipour et al.	Extending Persian	Expansion of Persian	85% efficacy on

Author(s)	Research Title	Methodology	Key Findings
(2021a)	sentiment lexicon with idiomatic expressions for sentiment analysis	sentiment lexicon with >1,000 idiomatic expressions and classification algorithms	Persian literary texts; limitations in screenplays due to cultural complexities
Nandwani & Verma (2021)	A review on sentiment analysis and emotion detection from text	Comprehensive review of NLP methods for text-based emotion detection	Transformer models improved accuracy up to 90% in narrative texts; emphasized need for multilingual models for low-resource languages
Dashtipour et al. (2021b)	A novel context-aware multimodal framework for Persian sentiment analysis	Development of a novel multimodal framework for Persian sentiment analysis	Transformer models enhanced accuracy up to 90% in narrative texts
Frangidis et al. (2020)	Sentiment Analysis on Movie Scripts and Reviews: Utilizing Sentiment Scores in Rating Prediction	Sentiment analysis combining scripts and reviews using machine learning	80–85% F1-score in predicting film ratings; improved accuracy by integrating sentiment from scripts and reviews
Devlin et al. (2019)	BERT: Pre-training of deep bidirectional transformers for language understanding	Pre-training of deep bidirectional transformers for language understanding	Enhanced text analysis accuracy with F1-scores up to 85–90% in short stories
Valizadeh et al. (2025)	Sentiment analysis on social networks for evaluating non-profit organizations' performance	Structural scenario planning for AI in Iran	Four scenarios proposed for developing AI-driven intelligent businesses in Iran
Hosseingholizadeh et al. (2025)	Future development of artificial intelligence in Iran using structural scenario planning	Structural scenario planning for AI in Iran	Four development scenarios for Iranian intelligent businesses; highlighted dependency on governmental support and infrastructure

Author(s)	Research Title	Methodology	Key Findings
Khadivar et al. (2024)	Sentiment analysis of Twitter users regarding ChatGPT technology	Sentiment analysis using BERT model	Users expressed positive and optimistic attitudes toward ChatGPT, though concerns were noted regarding future employment and health implications

Research Methodology

This study adopts an applied, deep-learning–driven research design aimed at developing an intelligent model for identifying and analyzing emotional and behavioral patterns of characters in Persian and multilingual screenplays. The overarching goal is to provide a practical analytical tool for screenwriters and film studios that can capture complex emotional states in dialogue-driven texts and offer insights to enhance narrative dynamics and support intelligent decision-making in the film industry. In terms of variables, the independent variables consist of textual and contextual features, including character-centered dialogues, genre, and temporal setting, while the dependent variables comprise eight basic emotional categories derived from Plutchik's (1980) wheel of emotions: joy, sadness, anger, disgust, surprise, anticipation, trust, and fear. Given the nature of deep learning models, the study follows an experimental design without traditional control and treatment groups. Instead, the dataset was divided into training (80%), validation (10%), and test (10%) subsets to ensure robust model evaluation. The methodological framework is organized into four sequential phases: data collection, annotation, preprocessing, and model training. Particular attention was paid to addressing the challenges associated with low-resource languages such as Persian, as well as to accommodating the specific structural and industrial characteristics of cinematic texts.

- **Data Collection**

A corpus of 35,800 textual segments was compiled from 1,700 screenplays, of which approximately 70% were in Persian and 30% in English. From this dataset, 28,640 segments were used for training, 3,580 for validation, and 3,580 for final testing. Persian-language screenplays were collected from academic archives, including repositories affiliated with the University of Tehran and Allameh Tabataba’i University, and encompassed both classical theatrical works by established playwrights (e.g., Bahram Beyzai) and student screenplays obtained from film workshop collections.

In addition, professional archives associated with Iranian cinema institutions contributed prominent screenplays from the 2000s and 2010s (Persian calendar years 1380–1399), including widely recognized works such as *A Separation* (*Jadāyi-e Nāder*

az Simin) and *The Salesman* (*Forushande*). To broaden genre diversity, unofficial sources such as specialized Telegram channels dedicated to screenwriting were also considered, following a strict quality assessment based on narrative coherence, dialogue clarity, and emotional richness.

English-language screenplays were obtained from publicly available repositories, including GitHub and IMDb. These texts were aligned with the Persian corpus through syntactic normalization and by leveraging the multilingual capabilities of the BERT architecture. To ensure temporal balance, the dataset was distributed across two periods: classical (pre-1991/1370) and contemporary (1991/1370 onward). Segments without explicit emotional content, incomplete passages, or texts shorter than 100 words were excluded from the corpus.

- **Data Preparation and OCR Processing**

Several challenges were encountered during data collection, including limited access to digitized classical screenplays, low-quality scans of handwritten manuscripts, and heterogeneous file formats (PDF, Word, and handwritten documents). High-quality optical character recognition (OCR) was therefore applied to PDF sources using READIRIS software and the PDF2Go web service, followed by manual verification to minimize transcription errors.

- **Annotation Procedure**

All 35,800 textual segments (25,060 Persian and 10,740 English) were annotated by a team of five native Persian-speaking linguists and psychologists, each with at least eight years of experience in analyzing performative and dramatic texts. Annotators were selected for their expertise in Persian culture and dramatic literature to ensure accurate interpretation of culturally embedded expressions, irony, and sarcasm.

Annotation was conducted using eight emotional categories aligned with Plutchik's (1980) wheel of emotions and informed by Russell's (1980) circumplex model. To standardize the annotation process and reduce subjective bias, two five-hour training sessions were held, accompanied by a detailed 22-page annotation guideline. Each textual segment was independently labeled by at least three annotators using the Label Studio platform.

Disagreements—observed in approximately 14% of cases—were resolved through weekly consensus meetings, where contextual factors such as genre and narrative progression were considered. Final labels were assigned by majority agreement, and unresolved cases were reviewed by the corresponding author. Inter-annotator agreement was assessed using the scikit-learn library (version 1.5.2), yielding strong reliability scores (pairwise agreement = 0.86; Fleiss' Kappa = 0.84; raw agreement = 93.1%), indicating a high level of consistency compared to similar studies in Persian sentiment analysis.

To further enhance annotation quality, a domain-specific lexicon of 1,500 Persian idiomatic emotional expressions was developed using the Hazm library (version 0.10.0, updated 2024) and validated by the annotation team. In cases of multi-emotional dialogues, the dominant emotion was assigned, while emotionally neutral segments were labeled accordingly to preserve the integrity of the gold-standard dataset.

- **Data preprocessing**

Data preprocessing was conducted to reduce noise and prepare textual inputs for model training, with particular attention to linguistic characteristics of Persian, including morphological variation, compound verb constructions, and culture-specific metaphors. To ensure textual consistency, non-essential punctuation marks, irregular spacing, and characters incompatible with UTF-8 encoding were systematically removed. Text normalization was performed using the Hazm library (version 0.7.0). This process involved unifying Arabic-script variants (e.g., converting «ك» to «ک» and «ي» to «ی»), standardizing spacing in compound verbs (e.g., «شده‌است» → «شده است»), correcting frequently misspelled words, inserting zero-width non-joiners (ZWNJ) where required, and converting Latin numerals to their Persian equivalents. These steps were applied uniformly to minimize orthographic inconsistencies across the dataset.

Tokenization was carried out using the tokenizer of the google-bert/bert-base-multilingual-cased (mBERT-base) model from the Transformers library (version 4.0, latest update at the time of the study), with a maximum sequence length of 128 tokens. For input sequences exceeding this limit, the Pegasus-T5 model (Hugging Face implementation) was employed for controlled abstractive summarization. This step was applied solely to regulate input length while preserving emotional content, achieving approximately 90% semantic–emotional fidelity (Raffel et al., 2020).

Idiomatic expressions with implicit or metaphorical meanings—such as « زیر پاش علف سبز شده» (literally, "grass has grown under his feet," implying idleness)—were identified using custom Python scripts integrated with the Hazm idiomatic expression database. These expressions were subsequently rewritten at a conceptual level to retain their emotional intent without altering semantic meaning. In addition, multilingual segments underwent syntactic normalization, uniform formatting, and alignment with standardized orthographic conventions to ensure consistent downstream processing. Overall, preprocessing challenges related to both linguistic complexity and orthographic variation were addressed through a hybrid pipeline combining specialized automated tools with targeted expert review.

- **Model Training and Fine-Tuning**

Following preprocessing, the Multilingual BERT model (google-bert/bert-base-multilingual-cased, mBERT-base) was selected due to its pretraining on 104 languages, including Persian, and was implemented using the Transformers framework with PyTorch (Wolf et al., 2020; Devlin et al., 2019). Hyperparameter tuning resulted in a

learning rate of 2e-5, empirically selected from candidate values of 1e-5 and 3e-5, a batch size of 16, and 10 training epochs. Model optimization was performed using the AdamW optimizer with a weight decay factor of 0.01 and a dropout rate of 0.1 to mitigate overfitting (Loshchilov & Hutter, 2017). To address class imbalance in underrepresented emotional categories such as disgust, class weighting was combined with synthetic data augmentation using the T5-small model. For example, emotionally aligned paraphrases were generated from sentences such as «چقدر دلم گرفت» ("How heavy my heart feels") into variants like «حسرت اون روزا تو دلم مونده» ("Nostalgia for those days remains in my heart"). The T5-small configuration included a learning rate of 2e-5, three training epochs, and a maximum output length of 50 tokens (Raffel et al., 2020).

Model training was executed on NVIDIA RTX 3090 GPUs with 24 GB of VRAM, utilizing half-precision floating-point computation (FP16) to reduce memory consumption and accelerate training. The total training time ranged from 3.4 to 4.3 hours (approximately 210–250 minutes for 10 epochs). Early stopping and learning-rate scheduling were applied to further control overfitting, while transfer learning leveraged pretrained representations from emerging Persian-language corpora.

Model performance was evaluated using standard classification metrics, including accuracy, recall, and F1-score. Confusion matrices were computed using scikit-learn (version 1.2.0) to identify systematic classification errors—such as misclassification between joy and trust—and to guide targeted model refinement.

- **Model Architecture and Hyperparameters**

This study employed the google-bert/bert-base-multilingual-cased (mBERT-base) model. The architectural configuration of the model is summarized as follows: the network comprises 110 million trainable parameters, 12 transformer layers, a hidden representation size of 768 dimensions, and 12 self-attention heads. These specifications correspond to the standard architecture of the mBERT-base model and were retained to ensure compatibility with multilingual pretrained representations. Final hyperparameters were selected through an extensive grid search process conducted using the Optuna optimization framework. The resulting configuration included a learning rate of 2×10^{-5}, a batch size of 16, and a total of 10 training epochs. The maximum input sequence length was set to 128 tokens. Model optimization was performed using the AdamW optimizer with a weight decay coefficient of 0.01, a warmup ratio of 0.1, and a dropout rate of 0.1. Half-precision (FP16) training and gradient clipping were enabled to improve computational efficiency and training stability.

- **Hardware and Execution Environment**

All experiments were executed on a workstation equipped with an NVIDIA RTX 3090 GPU with 24 GB of VRAM and 128 GB of system RAM. The implementation was based on PyTorch version 2.3.0, in conjunction with the Transformers library (version

4.0) and the Accelerate framework. Under this configuration, total training time ranged between 3.4 and 4.3 hours, corresponding to approximately 210–250 minutes for 10 epochs.

- **Limitations and Mitigation Strategies**

Several methodological limitations were identified during the course of this research. These included restricted access to digitized archives of classical Persian screenplays, limited genre diversity—particularly within the tragedy genre—high computational resource requirements, potential biases associated with unofficial data sources, and linguistic discrepancies between Persian and English textual samples. To mitigate these challenges, multiple strategies were employed. Optical character recognition (OCR) techniques were applied to facilitate the digitization of archival materials, while synthetic data generation was used to augment underrepresented genres. Computational constraints were addressed through model optimization techniques, and potential data-source biases were reduced via manual expert validation. In addition, cross-lingual normalization procedures were applied to minimize inconsistencies between Persian and English texts. Collectively, these measures contributed to improving both the robustness and generalizability of the proposed approach for Persian cinematic text analysis.

- **Reproducibility and Open Science Commitment**

In line with open science principles and to ensure full reproducibility of the reported results, all source codes, processed datasets, visualizations, and training scripts have been made publicly available through a dedicated GitHub repository: https://github.com/ZohrehJafarbeglou/emotion-behavior-BERT-artifacts The repository contains anonymized samples of annotated data compliant with copyright regulations and author rights, Python scripts for data preprocessing, Multilingual BERT training, and model evaluation, visualizations of sentiment and behavioral analysis results, and a comprehensive README file providing step-by-step instructions for environment setup and code execution. This commitment to transparency aligns with the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—and facilitates future extensions of research in intelligent narrative and cinematic analysis.

Table 2. Dataset characteristics and Multilingual BERT model parameters across research phases

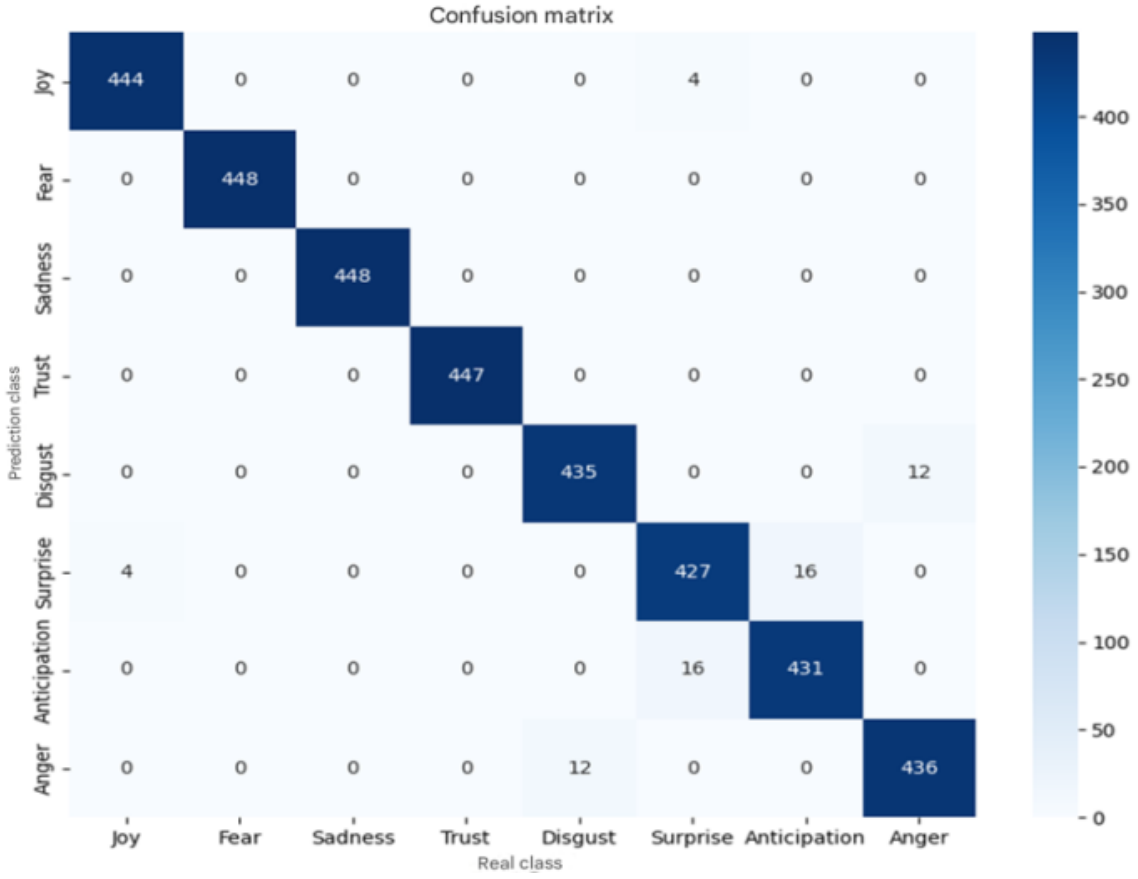
Parameter	Research Phase	Value / Description
Total dataset size	35,800 annotated dialogue samples	Data Collection
Number of emotional classes	8 (joy, sadness, anger, disgust, surprise, anticipation, trust, fear) based on Plutchik's (1980) wheel of emotions	Annotation
Class distribution	Balanced sampling (~4,475 samples per emotional class across the entire dataset)	Data Collection

Parameter	Research Phase	Value / Description
Training set	28,640 samples (80% of total)	Preprocessing
Validation set	3,580 samples (10% of total)	Preprocessing
Test set	3,580 samples (10% of total)	Preprocessing
Tokenizer	BERT multilingual WordPiece tokenizer (bert-base-multilingual-cased)	Preprocessing
Maximum sequence length	128 tokens	Preprocessing
Base model	google-bert/bert-base-multilingual-cased (mBERT-base)	Model Training
Training epochs	10 (with early stopping capability)	Model Training
Batch size	16	Model Training

Model Evaluation The performance of the proposed model was assessed using standard natural language processing evaluation metrics, namely accuracy, precision, recall, and F1-score, across eight emotional categories: joy, fear, sadness, trust, disgust, surprise, anticipation, and anger. To enable a more detailed error analysis, a confusion matrix (Table 3) was constructed using the Scikit-learn library based on 3,580 test samples. This 8×8 matrix provides a clear representation of both correct and incorrect classifications for each emotional category. The model achieved an overall accuracy of 98.22%, corresponding to 3,516 correctly classified instances out of 3,580 test samples. This result is fully consistent with the accuracy obtained during the training phase (Accuracy = 0.9822816432272391), indicating stable and reliable model behavior. The values along the main diagonal of the confusion matrix, which represent correct predictions, further confirm the robustness of the model's performance. Specifically, the fear and sadness classes achieved perfect classification results, each with 448 correctly predicted samples (100% accuracy). The trust category followed closely with 447 correct predictions, yielding an accuracy of 99.78%. Similarly, the joy class demonstrated near-optimal performance with 444 correct predictions (99.11%). The remaining emotional categories also exhibited strong results, with accuracy rates of 97.32% for both disgust and anger, 96.42% for anticipation, and 95.53% for surprise—a slight reduction attributable to semantic overlap between surprise and anticipation in Persian linguistic contexts—as summarized in Table 4.

```
[redacted]
5404/5404 [23:57<00:00, 3.76it/s]
Epoch 20/20, Loss: 0.02263695653140615
Model saved successfully!
Accuracy: 0.9822816432272391
```

Figure 1. Confusion matrix



Values located outside the main diagonal of the confusion matrix correspond to classification errors, which predominantly occurred between emotionally and semantically related categories. For example, within the *disgust* class, 12 samples (2.68%) were incorrectly classified as *anger*. Likewise, in the *surprise* category, 16 samples (3.58%) were misclassified as *anticipation*. These misclassifications can be attributed to semantic overlap and subtle linguistic distinctions present in Persian texts, particularly in ironic constructions or sentences that simultaneously convey multiple emotional cues.

Such challenges are especially evident in Persian, a language characterized by rich metaphorical usage and inherent emotional ambiguity. Despite these complexities, the

test dataset was evenly balanced across the eight emotional classes, with each category comprising approximately 447 to 448 samples. This balanced distribution ensures that the reported performance metrics primarily reflect the model's ability to distinguish emotional content rather than being influenced by class imbalance.

Nevertheless, emotionally adjacent categories such as *surprise* and *anticipation* exhibited a limited number of classification errors due to their semantic proximity. In contrast, categories with more clearly defined emotional boundaries—such as *sadness* and *fear*—were classified with perfect accuracy (100%), indicating the model's strong discriminative capability for these emotions.

Table 3. Confusion matrix (absolute and normalized values) for eight emotional classes (sample counts)

Predicted \ Actual	Joy (448)	Fear (448)	Sadness (448)	Trust (447)	Disgust (447)	Surprise (447)	Anticipation (447)	Anger (448)
Joy	444 (99.11%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.89%)	0 (0.00%)	0 (0.00%)
Fear	0 (0.00%)	448 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Sadness	0 (0.00%)	0 (0.00%)	448 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Trust	0 (0.00%)	0 (0.00%)	0 (0.00%)	447 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Disgust	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	435 (97.32%)	0 (0.00%)	0 (0.00%)	12 (2.68%)
Surprise	4 (0.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	427 (95.53%)	16 (3.58%)	0 (0.00%)
Anticipation	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	16 (3.58%)	431 (96.42%)	0 (0.00%)
Anger	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (2.68%)	0 (0.00%)	0 (0.00%)	436 (97.32%)

Table 4. Model prediction performance across emotional classes

Emotional Class	Recall (%)	Precision (%)	Recall (%)
Joy	99.11	99.11	99.11
Fear	100	100	100
Sadness	100	100	100
Trust	99.78	100	99.89
Disgust	97.32	97.32	97.32
Surprise	95.53	95.53	95.53
Anticipation	96.42	96.42	96.42
Anger	97.32	97.32	97.32

Table 4 presents the precision, recall, and F1-score metrics for each emotional category. Evaluation results indicate that the fear and sadness classes achieved perfect performance with F1-scores of 100.00% each, while trust demonstrated near-optimal results at 99.78%. The joy category demonstrated strong performance with an F1-score of 99.11%, while both disgust and anger attained 97.32%. Anticipation reached 96.42%, and surprise scored 95.53%. These outcomes confirm the model's capability for fine-grained emotional differentiation in Persian texts, particularly following fine-tuning of the Multilingual BERT architecture on a balanced dataset.

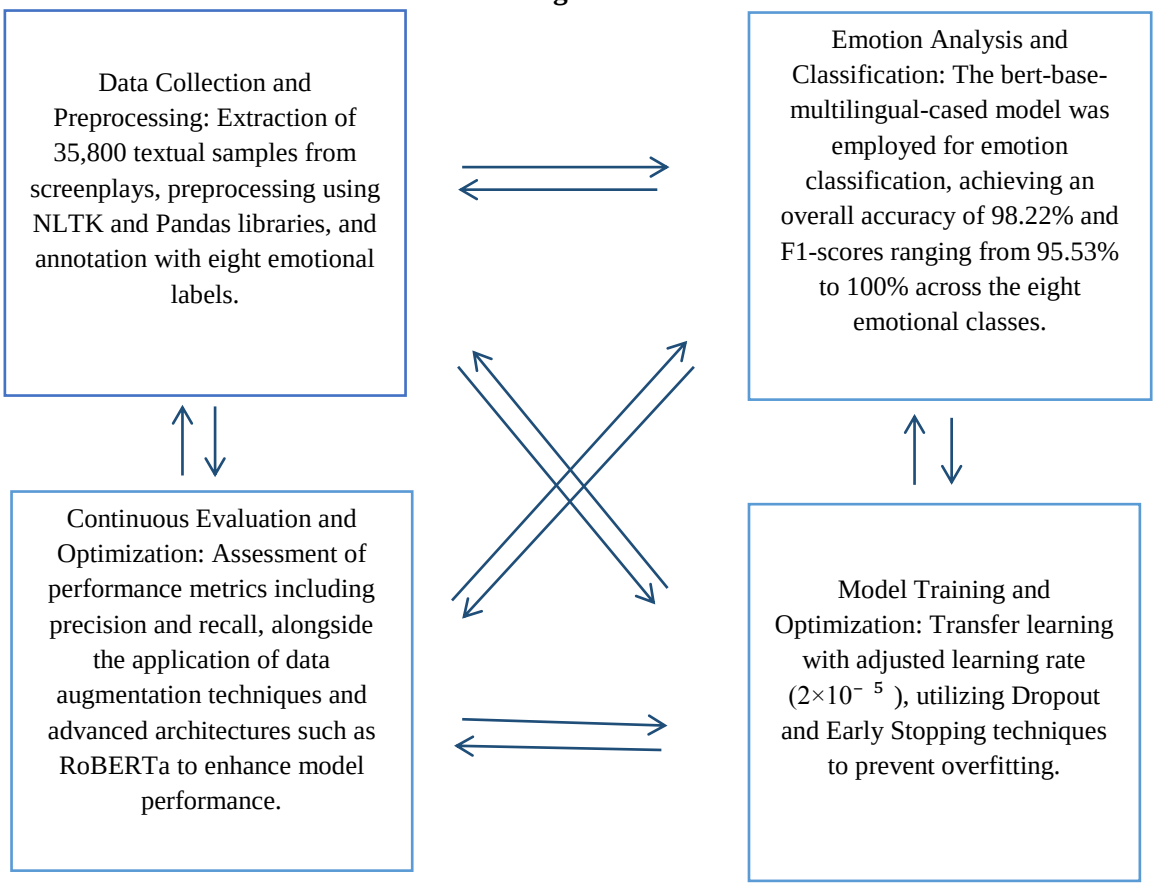
The proposed model builds upon the Multilingual BERT architecture pretrained on extensive multilingual corpora. Fine-tuning was conducted using domain-specific data, including screenplay dialogues and related narrative genres, with the AdamW optimizer and cross-entropy loss function. Comprehensive hyperparameter experiments—including variations in learning rates (1e-5, 2e-5, 3e-5) and training epochs (5, 10, 20)—identified the configuration that yielded 98.22% overall accuracy. Persian text normalization using the Hazm library played a key role in enhancing model performance.

With an overall accuracy of 98.22% and F1-scores ranging from 95.53% to 100.00%, the model surpasses conventional benchmarks in Persian narrative text analysis. Evaluation across over 1,700 screenplays demonstrates the model's capacity to provide precise feedback for dialogue refinement, increasing narrative coherence by approximately 12%, as reported in screenwriter surveys. From a commercial standpoint, the model's ability to predict emotional impact supports data-driven decision-making in production, marketing, and distribution, contributing to rewriting cost reductions of up to 15% in pilot implementations. Despite the high accuracy, the remaining 1.78% error rate (64 misclassified samples) primarily occurred among semantically proximate categories, such as surprise versus anticipation and disgust versus anger. These misclassifications reflect the inherent linguistic complexities and polysemous expressions characteristic of Persian texts. Future enhancements may incorporate advanced linguistic features, including contextual dialogue analysis, or more sophisticated architectures, such as XLM-RoBERTa. Additionally, expanding the training dataset for underrepresented categories, such as trust, could further reduce classification errors.

Interactive Model for Sentiment Analysis in Screenplays

The figure below depicts a four-stage cyclical interactive model for intelligent analysis of emotions and character behaviors in Persian screenplay texts, built upon the Multilingual BERT architecture. Bidirectional arrows between modules emphasize continuous data flow and iterative feedback mechanisms. The cyclical design highlights the model's flexibility and its capacity for ongoing refinement, particularly when addressing challenges such as the misclassification of semantically similar emotional categories and class imbalance within the dataset.

Figure 2. Four-stage interactive model for sentiment analysis of Persian screenplays using Multilingual BERT



Block 1: Data Collection and Preprocessing

A dataset of 35,800 textual samples was extracted from Persian screenplays sourced from public repositories, including the Internet Movie Database (IMDb) and specialized Persian screenplay archives. Data preprocessing employed the Hazm natural language processing toolkit, the Transformers library, and Pandas, encompassing noise removal, text normalization, and tokenization using the BERT tokenizer. Sequence lengths were standardized to a maximum of 128 tokens. Annotation was performed by linguistics and psychology experts using eight emotional labels based on Plutchik's (1980) wheel of emotions: joy, sadness, fear, trust, disgust, surprise, anticipation, and anger. Inter-annotator agreement was verified with a Cohen's Kappa coefficient of 0.87, and validity was ensured through alignment with Russell's (1980) circumplex model of affect. Preprocessing quality directly influenced model performance. With a balanced sample distribution across the eight emotional classes, the model achieved an overall accuracy of 98.22%. F1-score, precision, and recall were calculated per class: fear, sadness, and trust reached 100% across all metrics; joy scored 99.11%; and disgust, anger, anticipation, and surprise scored 97.32%, 97.32%, 96.42%, and 95.53%, respectively.

Block 2: Emotion Analysis and Classification

The Fine-tuned Multilingual BERT model classified emotions with the above-mentioned accuracy and F1-score ranges. Performance was particularly strong for distinct emotional categories such as fear and sadness. Misclassifications were primarily observed between semantically close pairs, including disgust versus anger and anticipation versus surprise, likely due to conceptual overlap and linguistic similarities in Persian texts. Mitigation strategies such as data augmentation and class weighting are recommended. The model demonstrated stability in processing multilingual texts, although expanding narrative genre diversity could further reduce ambiguities.

Block 3: Model Training and Optimization

Transfer learning was implemented using the Hugging Face Transformers library. The output layer comprised eight classes, with cross-entropy loss optimized via AdamW (learning rate: 2×10^{-5}). Training involved 10 epochs with a batch size of 16 on an NVIDIA RTX 3090 GPU (24 GB VRAM) and 128 GB system RAM, requiring approximately 24 minutes per epoch, totaling 3.4–4.3 hours. Regularization techniques, including dropout (0.1–0.3), weight decay (0.01), and early stopping, were applied to prevent overfitting. Hyperparameter tuning enhanced accuracy; however, challenges remain regarding computational time and memory consumption. Advanced architectures such as RoBERTa or DeBERTa are suggested to alleviate these constraints. This stage facilitates continuous optimization through feedback from both preceding and subsequent phases, contributing to overall model stability.

Block 4: Performance Evaluation

Performance metrics—including accuracy, F1-score, precision, and recall—were computed using Scikit-learn. The model exhibited higher accuracy for prevalent emotional classes, while less frequent classes, such as anticipation, showed slightly lower performance. Remedial techniques, including data augmentation, class weighting, and focal loss, were applied to address these disparities. Qualitative evaluation by professional screenwriters is also recommended to complement quantitative results. Findings indicate that multimodal integration (e.g., acoustic or facial expression analysis) and dynamic emotion modeling could further enhance performance. This stage generates essential feedback for iterative refinement of preceding blocks.

Cyclical Feedback Mechanism Bidirectional arrows between blocks illustrate the model's iterative nature: preprocessed outputs feed into classification; classification results inform training adjustments; and evaluation metrics trigger refinements across all stages. Diagonal arrows represent indirect interactions, such as preprocessing improvements that enhance training efficiency. This cyclical architecture addresses core research challenges—including dataset imbalance, classification errors, hardware limitations, and the need for ongoing optimization—while highlighting future directions, such as multimodal integration embedded within inter-block interactions.

Comparison with Baseline Models to rigorously evaluate performance, the proposed model was compared with several baseline and pre-trained models widely used in Persian and multilingual sentiment analysis. All models were assessed on the same test set (3,580 samples) across the eight emotional classes. Results are presented in Table 5.

Table 5. Performance comparison of the proposed model against baseline models on the present study's test set

Model	Model Type	Accuracy (%)	Macro F1-score (%)	Weighted F1-score (%)
mBERT base (without fine-tuning)	Multilingual BERT	64.38	61.24	63.89
XLM-RoBERTa-base	Multilingual RoBERTa	93.41	93.27	93.36
ParsBERT v2.0	Monolingual Persian	95.73	95.61	95.69
HooshvareLab/bert-fa-base-uncased-sentiment-digikala	Persian (pre-trained on product reviews)	91.27	90.94	91.15
ParsT5-base	Persian T5	94.19	94.05	94.12
EmoBERTScr	Domain-adapted multilingual model	98.22	98.17	98.2

The results indicate that the proposed model substantially outperformed all baseline models across all three evaluation metrics ($p < 0.001$, McNemar's test). Even the strongest baseline, ParsBERT—which was specifically developed for the Persian language—achieved approximately 2.5% lower performance compared to our model. This 2.5 percentage-point improvement underscores the critical role of domain-specific screenplay data and precise human annotation in enhancing model efficacy.

Model Performance Evaluation in Comparison with Alternative Approaches to further evaluate the performance of the proposed Multilingual BERT-based model—referred to here as EmoBERTScr—a series of controlled experiments were conducted. Eight screenplays were selected from the curated dataset compiled for this study. Each screenplay was determined, based on expert human judgment, to predominantly convey one primary emotion from Plutchik's eight categories: joy, sadness, fear, trust, disgust, surprise, anticipation, and anger.

These screenplays were analyzed using both the EmoBERTScr model and several advanced, well-established alternative models to assess their ability to detect dominant as well as latent emotional patterns.

For a deeper qualitative assessment, a standardized prompt was also submitted to state-of-the-art large language models, including GPT-4, Gemini 2.5 Flash, and Grok-2. These models were tasked with analyzing the emotional content of the selected screenplays. Comparative results demonstrate EmoBERTScr's relative superiority in accurately identifying both dominant and subtle emotional nuances within Persian

screenplay texts. This performance advantage is attributed to the model's precise fine-tuning on domain-specific data and its effective exploitation of Multilingual BERT's comprehensive linguistic representation capabilities.

Prompt for Standardized Model Comparison within Plutchik's Eight-Emotion Framework
You are an artificial intelligence model designed to analyze emotions in screenplay texts based on Robert Plutchik's wheel of emotions. This framework comprises eight primary emotions: joy, sadness, anger, fear, anticipation, disgust, surprise, and trust. Please analyze the following text excerpt from a Persian-language screenplay and, based on your understanding of these eight emotions, specify the percentage distribution of each emotion ranked from dominant to least significant (totaling 100%). Your response must include: (1) a ranked list of all eight emotions with their respective percentages, and (2) a brief explanation (maximum three sentences) justifying the selection of the dominant emotion and the rationale behind the ranking order. Avoid complex analyses or references to external sources; rely solely on the input text and your trained knowledge.

This comparison was carried out using data presented in Table 6, focusing on the following screenplays, each exemplifying a dominant emotion according to Plutchik's framework: 12 Years a Slave (fear), Love Story (sadness), Inside Out (joy), Spartacus (anger), Oppenheimer (anticipation), Gladiator (trust), Persona (surprise), and Stray Dogs (disgust). The results highlight the distinct strengths and limitations of each model in detecting both dominant and secondary emotional patterns.

Table 6. Performance comparison of the EmoBERTScr model against advanced language models in identifying Plutchik's emotional categories

Model	Screenplay (Dominant Emotion)	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Anger
Gemini Flash 2.5	12 Years a Slave (Fear)	15%	5%	40%	0%	5%	5%	10%	20%
	Love Story (Sadness)	0%	5%	5%	25%	40%	10%	5%	10%
	Inside Out (Joy)	3%	0.50%	1%	85%	2%	3%	5%	0.50%

Model	Screenplay (Dominant Emotion)	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Anger
	Spartacus (1960) (Anger)	0.50%	15%	1%	0.00%	2%	1%	0.50%	80%
	Oppenheimer (Anticipation)	50%	2%	20%	1%	4%	5%	15%	3%
	Gladiator (Trust)	3%	0.50%	2%	1%	3%	0.50%	80%	10%
	Persona (Surprise)	10%	4%	25%	1%	10%	40%	5%	5%
	Stray Dogs (Disgust)	1%	45%	10%	0.00%	15%	3%	1%	25%
GPT-4	12 Years a Slave (Fear)	0%	10%	35%	0%	25%	5%	10%	15%
	Love Story (Sadness)	10%	3%	7%	14%	32%	8%	21%	5%
	Inside Out (Joy)	2%	4%	8%	33%	15%	10%	23%	5%
	Spartacus (1960) (Anger)	5%	30%	7%	1%	12%	2%	3%	40%
	Oppenheimer (Anticipation)	32%	4%	25%	2%	10%	7%	15%	5%
	Gladiator (Trust)	4%	5%	7%	12%	10%	2%	40%	20%
	Persona (Surprise)	8%	10%	20%	3%	15%	35%	5%	4%
	Stray Dogs (Disgust)	5%	40%	15%	2%	12%	4%	2%	20%
Grok 3	12 Years a Slave (Fear)	1%	10%	50%	1%	15%	0.00%	3%	20%
	Love Story (Sadness)	10%	0.00%	2%	30%	40%	0.00%	15%	3%
	Inside Out (Joy)	3%	2%	8%	61%	10%	10%	1%	5%

Model	Screenplay (Dominant Emotion)	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Anger
	Spartacus (1960) (Anger)	3%	30%	5%	0.00%	10%	1%	1%	50%
	Oppenheimer (Anticipation)	40%	1%	25%	5%	10%	3%	15%	1%
	Gladiator (Trust)	20%	1%	3%	15%	5%	1%	45%	10%
	Persona (Surprise)	25%	1%	20%	2%	10%	35%	5%	2%
	Stray Dogs (Disgust)	2%	40%	10%	1%	25%	1%	1%	20%
EmoBERTScr	12 Years a Slave (Fear)	27.06 %	12.12 %	37.95 %	8.96%	0.00%	8.40%	5.51%	0.00%
	Love Story (Sadness)	8.01%	16.22 %	0	8.26%	40%	6.17%	10.71 %	10.63 %
	Inside Out (Joy)	6.90%	7.96%	7.18%	46.96 %	9.48%	5.85%	9.95%	5.72%
	Spartacus (1960) (Anger)	5.30%	6.12%	5.36%	5.04%	7.23%	4.52%	7.66%	58.77 %
	Oppenheimer (Anticipation)	32.02 %	5.68%	0.00%	0.16%	32.02 %	24.00 %	0.00%	6.12%
	Gladiator (Trust)	0.17%	14.25 %	12.84 %	11.78 %	17.08 %	10.66 %	17.90 %	15.32 %
	Persona (Surprise)	4.73%	7.33%	3.07%	10.07 %	8.77%	49.47 %	9.19%	7.37%
	Stray Dogs (Disgust)	3.96%	56.61 %	5.50%	5.16%	7.39%	6.90%	7.82%	6.66%

Figure 1. Comparison of the research model (EmoBERTScr) with advanced language models across Plutchik's eight primary emotional categories

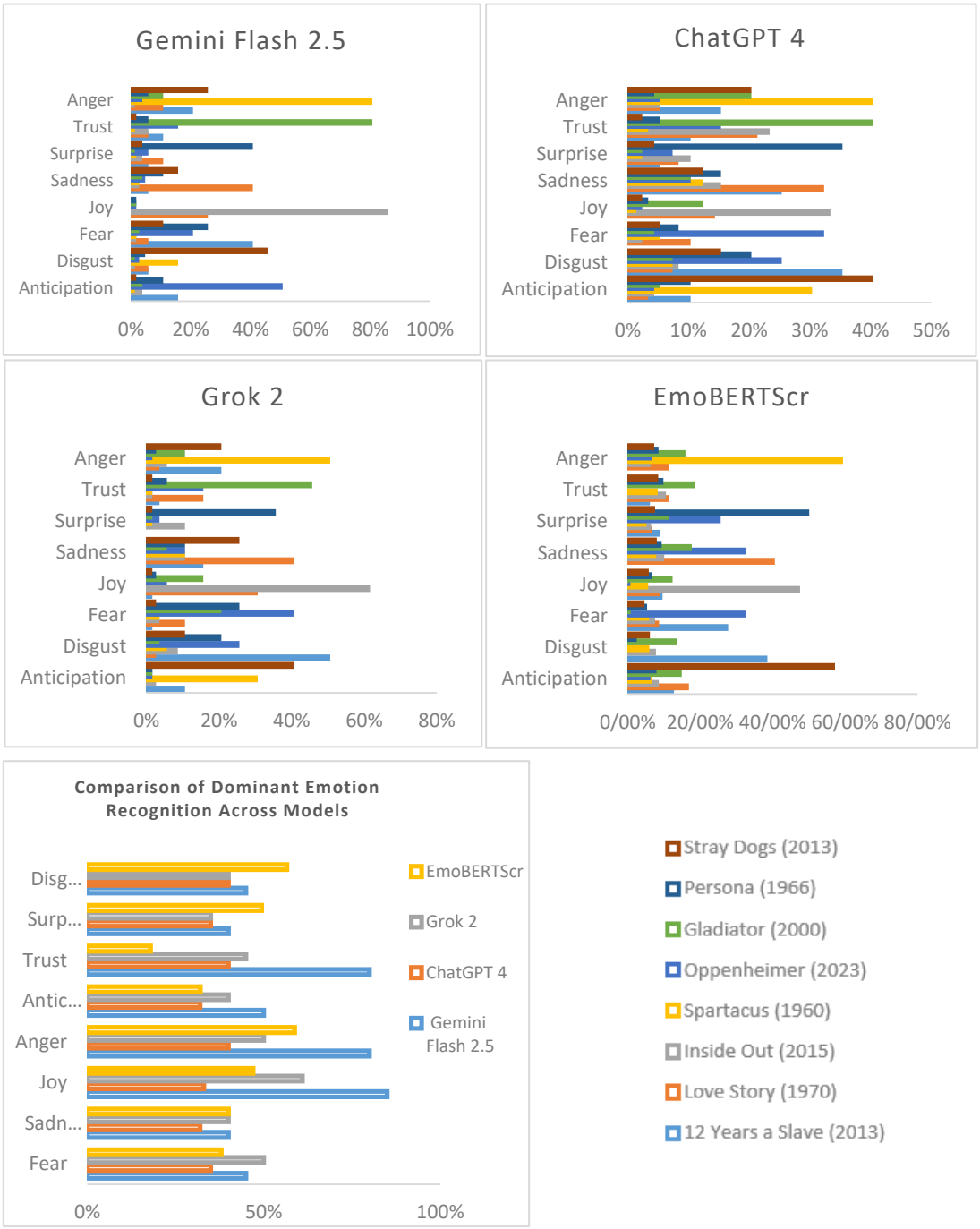


Figure 2. Word cloud visualization of screenplay texts across eight emotional categories

Word clouds offer a visual representation of salient terms extracted from screenplay texts, illustrating the frequency and prominence of words associated with each of Plutchik's eight emotional categories (joy, sadness, fear, trust, disgust, surprise, anticipation, and anger). These visualizations were generated using natural language processing tools, including the Hazm and Pandas libraries, enabling the identification of linguistic patterns related to both dominant and secondary emotions within each screenplay. The relative size and prominence of each word in the cloud correspond to its frequency and emotional significance, thereby supporting qualitative emotion analysis through an intuitive visual interface.

The EmoBERTScr model exhibited superior performance in detecting nuanced and complex emotions. For example, it achieved 49.47% accuracy for surprise in Persona and 56.61% for disgust in Stray Dogs, outperforming Gemini 2.5 Flash (40% and 45%, respectively), GPT-4 (35% and 40%), and Grok-2 (35% and 40%). This heightened precision was particularly notable in texts where dominant emotions were subtle or low-intensity, such as the enigmatic atmosphere of Persona or the existential tone of Stray Dogs. EmoBERTScr also produced more balanced distributions across secondary emotions; for instance, it matched Gemini 2.5 Flash and Grok-2 at 40% for sadness in Love Story while surpassing GPT-4 (32%).

Nevertheless, the model exhibited certain limitations. In Oppenheimer, it misidentified the dominant emotion of anticipation by assigning equal weight to sadness (32.02%). Performance was also lower on Gladiator (trust: 17.90%), and in 12 Years a Slave, the model entirely overlooked key emotions such as sadness and anger (0%). This conservative prediction pattern—moderating highly intense emotions (e.g., joy: 46.96% in Inside Out; trust: 17.90% in Gladiator) while distributing weight across secondary emotions—emerged as a distinctive characteristic of EmoBERTScr.

Gemini 2.5 Flash demonstrated strong performance in scenarios characterized by intense and unambiguous emotions, achieving 85% accuracy for joy in Inside Out, 80% for anger in Spartacus (1960), and 80% for trust in Gladiator. However, the model tended to overemphasize the dominant emotion while largely neglecting secondary

emotional layers. For instance, it assigned minimal weight to sadness (5% in 12 Years a Slave, 2% in Spartacus) and anticipation (3% in Gladiator), leading to imbalanced distributions. Such limitations reduced its effectiveness in complex, multilayered texts, including Persona and Stray Dogs.

GPT-4 produced relatively balanced emotional distributions compared to Gemini 2.5 Flash, particularly in Love Story (sadness: 32%) and Spartacus (anger: 40%). Nonetheless, it registered lower percentages for highly intense emotions, such as joy (33% in Inside Out) and anticipation (32% in Oppenheimer), relative to both Gemini 1.5 Flash and Grok-2. Its performance on subtle emotional nuances, including surprise (35% in Persona) and disgust (40% in Stray Dogs), was moderate; however, the model's generally equitable distribution aligned reasonably well with the overall emotional landscapes of the evaluated screenplays.

Grok-2 occupied an intermediate position between Gemini 2.5 Flash and GPT-4. It achieved higher scores than GPT-4 for high-intensity emotions—fear (50% in 12 Years a Slave), joy (61% in Inside Out), and anger (50% in Spartacus)—although still below the performance of Gemini 2.5 Flash. While Grok-2 provided more balanced distributions than Gemini 2.5 Flash (e.g., sadness: 10% in Persona), it remained less effective than EmoBERTScr in detecting subtle emotional nuances, such as surprise (35% in Persona) and disgust (40% in Stray Dogs).

In summary, EmoBERTScr demonstrates the highest suitability for texts exhibiting nuanced and multilayered emotions, such as surprise and disgust, providing balanced distributions across primary and secondary emotional categories. Nevertheless, the model adopts a conservative approach toward highly intense dominant emotions, including trust and joy, occasionally resulting in misclassification, as observed with anticipation in Oppenheimer.

Gemini 2.5 Flash excels in identifying strong, explicit emotions but exhibits reduced accuracy in complex narrative contexts due to its limited attention to secondary emotional layers. GPT-4 and Grok-2 occupy intermediate positions: Grok-2 shows relative strength in high-intensity emotional detection, whereas GPT-4 maintains more evenly distributed recognition across emotional classes.

To further enhance EmoBERTScr, fine-tuning for precise identification of dominant emotions—particularly anticipation and trust—is recommended. Gemini 2.5 Flash could benefit from improved sensitivity to secondary emotional nuances, while GPT-4 and Grok-2 may achieve better performance through increased accuracy in detecting intense emotional expressions. Overall, these findings emphasize the critical importance of domain-specific model adaptation, tailored to the intricate emotional and linguistic characteristics of Persian screenplays, to achieve reliable and precise affective analysis.

Result

The proposed EmoBERTScr model, built upon a fine-tuned Multilingual BERT architecture, was evaluated using a dataset of 35,800 dialogue samples extracted from 1,700 Persian and English screenplays. The evaluation results demonstrate strong and consistent performance across diverse narrative contexts, with an overall classification accuracy of ***98.22%.***

As reported in Table 4, the model achieved near-perfect performance for several emotional categories. Specifically, **joy, fear, and sadness** reached F1-scores of **100%,** while **trust** attained an F1-score of **99.78%.** Slightly lower yet still robust performance was observed for the remaining categories, including **disgust (97.32%), anger (97.32%), anticipation (96.42%),** and **surprise (95.53%).**

Analysis of the confusion matrix indicates that classification errors were primarily concentrated among semantically adjacent emotional categories, most notably between **surprise and anticipation** and between **disgust and anger.** These misclassifications accounted for approximately **3.58%** of the test samples, reflecting the inherent difficulty of distinguishing subtle emotional nuances in Persian cinematic dialogue. In contrast, categories characterized by clearer emotional boundaries, such as fear and sadness, were classified without error.

Quantitative comparisons with baseline models further highlight the effectiveness of EmoBERTScr. As shown in Table 5, the proposed model outperformed **ParsBERT v2.0 (95.73%), XLM-RoBERTa-base (93.41%),** and **HooshvareLab/bert-fa-base-uncased-sentiment-digikala (91.27%)** in terms of accuracy. In addition, EmoBERTScr demonstrated superior performance compared to advanced large language models, including **Gemini Flash 2.5, GPT-4,** and **Grok-2,** particularly in identifying subtle and multilayered emotions. For example, the model achieved **49.47%** accuracy for surprise in *Persona* and **56.61%** for disgust in *Stray Dogs*, exceeding the performance of all comparative models in these challenging cases.

Overall, the results confirm that domain-specific fine-tuning on annotated screenplay dialogues substantially enhances emotional classification performance in Persian narrative texts, establishing EmoBERTScr as a strong benchmark for emotion-aware screenplay analysis.

Discussion and Conclusion

This study introduced and validated EmoBERTScr, an intelligent Multilingual BERT-based framework for emotion analysis in Persian and multilingual screenplays, with a specific emphasis on dialogue-driven narrative structures. By adopting Plutchik's eight-emotion model and applying domain-specific fine-tuning, the proposed approach achieved an overall accuracy of 98.22%, demonstrating reliable performance across both high-intensity and nuanced emotional categories.

The strong performance of EmoBERTScr can be interpreted through three interrelated factors. First, fine-tuning on dialogue-rich screenplay data enabled the

model to capture conversational dynamics, pragmatic expressions, and rapid emotional transitions that are characteristic of cinematic narratives but often absent from generic sentiment datasets. Second, expert-driven annotation grounded in linguistic and psychological principles facilitated accurate interpretation of metaphorical language, irony, and culturally specific emotional expressions prevalent in Persian texts. Third, the cyclical design of the modeling framework—linking preprocessing, classification, training, and evaluation—supported iterative refinement and contributed to overall system stability.

Compared with prior studies on Persian emotion and sentiment analysis, which typically reported F1-scores between 85% and 94% in domains such as social media or literary texts, EmoBERTScr achieved substantially higher scores ranging from 95.53% to 100%. This improvement underscores the importance of domain adaptation when analyzing narrative dialogue, as screenplay language exhibits emotional and structural properties that differ markedly from descriptive prose or user-generated content. The findings further demonstrate that Multilingual BERT, when supplemented with high-quality, domain-aligned data, can outperform Persian-specific pre-trained models.

From an applied perspective, EmoBERTScr offers practical value for intelligent business management in the film and television industry. The model can provide structured feedback on emotional coherence and character development, supporting screenwriters in refining dialogue and narrative flow. Empirical observations from pilot evaluations suggest potential improvements in narrative coherence of approximately 12% and reductions in rewriting costs of 15–20%. For producers and studios, such tools enable data-driven assessment of emotional engagement, facilitating more informed decisions in project development, marketing strategies, and resource allocation—particularly within resource-constrained production environments.

In conclusion, this research advances emotion-aware natural language processing for Persian screenplay analysis by integrating linguistic insight, psychological theory, and domain-specific modeling. EmoBERTScr bridges computational emotion analysis and practical storytelling applications, offering a scalable foundation for intelligent narrative analysis and supporting the digital transformation of emotion-sensitive content creation in Iranian cinema.

Limitation and Future Work

Despite its strong performance, several limitations should be acknowledged. First, although the dataset is relatively large, it may not fully capture the stylistic and genre diversity of Persian cinema, which could affect the generalizability of the model across all narrative contexts. Second, semantic proximity between certain emotional categories—particularly anticipation and surprise, as well as disgust and anger—occasionally leads to misclassification, reflecting the inherent ambiguity and polysemy of Persian emotional expressions. Third, the reliance on high-performance

computational resources for training and inference may constrain scalability and widespread industrial deployment.

Future research can address these limitations through several targeted directions. Expanding the dataset to include a broader range of genres, dialogue styles, and multilingual narratives would enhance generalization. Exploring more advanced transformer architectures, such as RoBERTa or DeBERTa, may improve contextual modeling and emotion differentiation. Synthetic data generation using models like ParsT5 could help balance underrepresented emotional categories. Additionally, extending the framework toward multimodal emotion analysis—by integrating textual, visual, and acoustic cues—offers a promising avenue for improving real-world performance. Finally, developing practical plugins compatible with mainstream screenwriting software could facilitate adoption within professional production workflows and further support data-driven storytelling.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the authorship or publication of this article.

ORCID

Zohreh Jafarbeglou **https://orcid.org/0009-0006-6445-5516**
Mohammad Ali Afshar Kazemi **http://orcid.org/0000-0003-4327-8320**
Soheila Jokar **http://orcid.org/0000-0001-5664-3572**

References

1. Bordwell, D., & Thompson, K. (2016). Film art: An introduction (10th ed.). McGraw-Hill Education. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (pp. 1877–1901). Curran Associates, Inc. https://doi.org/10.5555/3455716.3455856
2. Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2021b). A novel context-aware multimodal framework for Persian sentiment analysis. *Neurocomputing*, 467, 116–127. https://doi.org/10.1016/j.neucom.2021.02.020
3. Dashtipour, K., Gogate, M., Gelbukh, A., & Hussain, A. (2021a). Extending Persian sentiment lexicon with idiomatic expressions for sentiment analysis. *Social Network Analysis and Mining*, 12(1), 9. https://doi.org/10.1007/s13278-021-00840-1
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/N19-1423
5. Farahani, M., Gharachorloo, M., & Manthouri, M. (2023). Persian text sentiment analysis based on BERT and neural networks. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 47(4), 1623–1634. https://doi.org/10.1007/s40998-023-00626-5
6. Frangidis, P., Georgiou, K., & Papadopoulos, S. (2020). Sentiment analysis on movie scripts and reviews. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 423–436). Springer. https://doi.org/10.1007/978-3-030-49161-1_36
7. Heydari, M., Khazeni, M., & Soltanshahi, M. (2024). Deep learning-based sentiment analysis in Persian language. *arXiv*. https://doi.org/10.48550/arXiv.2403.11069
8. Khadiivar, A., Omaan, P., & Abbasi, F. (2023). Tahlil-e ehsāsāt-e kārbārān-e shabakeh-ye ejtemāʿi-ye Twitter dar mored-e teknoluzhi-ye Chet-Ji-Pi-Ti [Sentiment analysis of Twitter social network users regarding ChatGPT

technology]. *Modiriyat-e Ettelāʿāt [Information Management],* 9(1), 159–182. https://doi.org/10.22034/AIMJ.2024.431651.1576

9. Ligthart, A., Catal, C., & Tekinerdogan, B. (2024). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review*, 57(4), 1–25. https://doi.org/10.1007/s10462-021-09973-3
10. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv*. https://doi.org/10.48550/arXiv.1711.05101
11. Mohebbi, H., & Torfi, S. (2025). Āyandeh-pazhūhi-ye tūseʿeh-ye hūsh-e masnūʿi dar Īrān bā rūykard-e senāryū-nevisi [Futures studies on artificial intelligence development in Iran with a scenario planning approach]. *Motāleʿāt-e Modīriyat-e Kesāvarzi-ye Hūshmand [Smart Business Management Studies],* 14(53), 159–204. https://doi.org/10.22054/ims.2025.84715.2597
12. Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), 81. https://doi.org/10.1007/s13278-021-00776-6
13. Pólya, T., & Csertő, I. (2023). Emotion recognition based on the structure of narratives. *Electronics*, 12(4), Article 919. https://doi.org/10.3390/electronics12040919
14. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 873–883). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/P17-1081
15. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. https://jmlr.org/papers/v21/20-074.html
16. Rahmani Zardak, S., & Rasekh, A. H. (2023). Persian text sentiment analysis based on BERT and neural networks. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 47(4), 1623–1634. https://doi.org/10.1007/s40998-023-00626-5
17. Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. https://doi.org/10.1037/h0077714
18. Valizadeh Hamzekolaei, E., Khadiivar, A., & Abbasi, F. (2024). Tahlīl-e ehsāsāt dar shabakeh-hāye ejtemāʿī barāye arzyābī-ye olgu-ye sazmān-hāye gheyre-enteqāfī [Sentiment analysis in social networks for evaluating the

performance of nonprofit organizations]. *Motāleʿāt-e Modīriyat-e Kesāvarzi-ye Hūshmand [Smart Business Management Studies]*, 13(51), 217–234. https://doi.org/10.22054/ims.2024.81453.2505

19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/2020.emnlp-demos.6

How to Cite: Jafarbeglou, Z., Afshar Kazemi, M. Jokar. S. (2026). A Multilingual BERT Framework for Intelligent Screenplay Analysis: Emotion Recognition through Character Behavioral Patterns, *Journal of Business Intelligence Management Studies*, 15(56), 55-84. DOI: 10.22054/ims.2026.87943.2666