Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A New Framework for Multilingual Sentiment Analysis in Natural Language Processing Using a Pre-trained BERT Model Supported by Multi-class Machine Learning Methods

Nazila Panahi

na.panahi@urmia.ac.ir

Mehdi Chehel Amirani

Research Article

Keywords: Natural Language Processing (NLP), Sentiment Analysis, BERT Model, Deep Learning (DL), Machine Learning (ML), Sentiment Analysis, Multilingual Analysis, Cyclostationary Analysis

Posted Date: June 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9734879/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A New Framework for Multilingual Sentiment Analysis in Natural Language Processing Using a Pre-trained BERT Model Supported by Multi-class Machine Learning Methods

Nazila Panahi[1][0000-0003-2846-4184] and Mehdi Chehel Amirani[2][0000-0002-5179-9831]

[1, 2] Independent Researcher (Personal Capacity), Urmia, Iran.

Nazila Panahi (corresponding author) : na.panahi@urmia.ac.ir

Abstract. Sentiment analysis, a central application of Natural Language Processing (NLP), identifies emotions and opinions in text. Accurate interpretation requires accounting for linguistic and cultural variations across languages. Considering text as a signal and given the cyclostationarity of most human-constructed signals, we sought to enhance sentiment analysis of human-written notes by extracting the most informative features, following our previous studies. From a signal-processing perspective, we examine the cyclostationarity of text embeddings to reveal hidden periodic patterns and describe their statistical properties in the frequency domain. Our approach combines advanced feature extraction and enhancement with classification and regression models to improve BERT's performance without modifying its architecture. Experiments on five benchmark datasets in four languages—English, Arabic, Brazilian Portuguese, and Turkish—and a mixed Arabic-English corpus showed up to a 28% accuracy improvement. Results confirm that human-generated text exhibits cyclostationary properties that can be leveraged for more effective sentiment modeling. This research represents the first comprehensive study of multilingual alignment in ML- and DL-based NLP sentiment analysis and demonstrates how signal-processing-based features can increase model robustness and cross-lingual accuracy.

Keywords: Natural Language Processing (NLP), Sentiment Analysis, BERT Model, Deep Learning (DL), Machine Learning (ML), Sentiment Analysis, Multilingual Analysis, Cyclostationary Analysis.

1 Introduction

Artificial intelligence has had a tremendous impact on every aspect of the world. It is widely used in various fields to achieve desired outcomes efficiently and accurately within a specific time frame [1]. Natural language processing (NLP) is a branch of artificial intelligence (AI) that concentrates on how computers and humans communicate using natural language. [2]. Sentiment analysis, a key application of NLP, is used to identify and extract the emotional tone or opinion expressed in text. It helps determine whether a text is positive, negative, or neutral. This technique is widely used in fields like marketing, customer service, healthcare, finance, and government decision-making, enabling organizations to assess public opinion, monitor brand reputation, and predict market trends [3]. However, the rapid expansion of the internet and consumers' active participation in sharing, commenting, and discussing on blogs, forums, social media platforms, and online shopping sites has made sentiment analysis a particularly dynamic study area [4].

The field's journey is filled with complexities, mainly due to the variability and subtlety of human emotions, combined with language's ever-changing and diverse nature. The challenges of linguistic diversity, contextual subtleties, and the evolution of online communication methods require advanced machine learning techniques capable of handling these complexities [5]. Understanding and interpreting human emotions through text improves human-computer interactions and offers valuable insights into public opinions and trends, enabling better decision-making across various fields [6].

As global interconnectedness increases, opinions and sentiments are more frequently expressed in various languages and dialects, sometimes within a single document [7]. Multilingual sentiment analysis has become a key research area in NLP, addressing the challenges of understanding opinionated text and the complexities of language diversity. These challenges become especially difficult in news sentiment analysis, social media monitoring, and other large-scale applications where texts are lengthy and contextually varied.

Unlike typical deep learning (DL) sentiment analysis studies that focus on enhancing the DL model itself, this research explores the signal-processing perspective on embeddings of a common DL model. The goal is to refine the features extracted by BERT, enhancing its sentiment analysis capabilities without modifying the core BERT architecture. The proposed approach is tested on five well-known datasets in four languages: Arabic, Arabic-English, Turkish, Brazilian Portuguese, and English. These languages represent diverse linguistic families and structures, making them ideal for a

comprehensive evaluation of the scheme and its performance, demonstrating its ability to handle multilingual processing. This is the first thorough review of multilingual alignment in ML-DL NLP Sentiment Analysis. The main contributions of this work include:

- The signal-processing perspective on embeddings
- Enhance sentiment analysis without changing the BERT architecture.
- Multilingual Sentiment Analysis by a unified model

The remainder of the paper is organized as follows: the next section reviews previously published studies, the third section explores the research methodology, the fourth section explains the presented scheme, the fifth section presents simulation results and discussions, and the last section concludes and provides suggestions for future work.

2 Previous Studies

This section presents a review of methods and the studied datasets. A historical overview and the evolution of sentiment analysis methods are discussed in [8]. The discussion starts with traditional machine learning and lexicon-based techniques, which formed the foundation for sentiment classification. Then it moves on to more advanced approaches such as deep learning, transfer learning, and hybrid models. This progression reflects the field's development toward more sophisticated and versatile models.

Traditional machine-learning methods have played a vital role in the initial phases of sentiment analysis. Techniques such as SVMs, Naïve Bayes classifiers, ensemble methods, and recursive neural tensor networks typically involve extracting features from text, like word frequencies or patterns, before applying classification algorithms to identify sentiment [9-12]. This foundational process sets the stage for more advanced techniques to follow. Hybrid approaches that combine Naïve Bayes and SVM leverage the benefits of both, emphasizing the importance of integrated feature selection and classification. The success of these traditional models largely depends on feature choices and the dataset structure.

Rule-based and lexicon-based methods are some of the earliest techniques in sentiment analysis [8]. These methods are known for their simplicity, interpretability, and efficiency. Rule-based approaches depend on predefined linguistic rules to classify sentiment [13], while lexicon-based techniques use dictionaries of sentiment-labeled words to assign sentiment scores [14]. Although both methods are computationally efficient and perform well in domains with limited training data, they are inherently limited by their dependence on static lexicons and predefined rules. These approaches often struggle to adapt to new datasets and have difficulty capturing the semantic nuances of natural language, such as idiomatic expressions and context-dependent sentiment. Another issue with lexicon-based methods is the need to maintain and update sentiment lexicons to include domain-specific vocabulary and accommodate changes in language usage.

Deep learning models have significantly enhanced sentiment classification accuracy by directly learning hierarchical and semantic features from data [8]. However, they need large labeled datasets and substantial computational resources, and are generally non-transparent in decision-making. Transfer learning techniques, especially those employing pre-trained transformer models such as BERT and RoBERTa, have achieved state-of-the-art performance, offering robustness across domains with minimal supervision. Their main limitations include high pretraining costs and limited interpretability. Convolutional Neural Networks (CNNs) are effective at capturing local semantic patterns [15,16], while Recurrent Neural Networks (RNNs), especially Long Short-Term Memory networks (LSTMs), excel at modeling sequential dependencies in language [17]. Combining CNNs, LSTMs, and attention mechanisms in hybrid models has further pushed performance boundaries [18]. Recent progress involves incorporating Transformer-based models like BERT, which have set new benchmarks in transfer learning for sentiment analysis [19]. These trends point to increasing use of hybrid architectures and attention mechanisms to enhance sentiment classification.

Transfer learning has become a key method in sentiment analysis, enabling models to use knowledge from one domain and apply it to another, often with limited labeled data [8]. This method is particularly useful when sentiment varies significantly across different domains or languages. Typically built on deep neural networks, transfer learning employs pre-trained models to address sentiment classification challenges across diverse fields effectively. It involves pre-training a neural network on a large source dataset and then fine-tuning it on a smaller, domain-specific target dataset. This allows the model to adapt to new environments with minimal additional training by utilizing the neural networks' ability to identify complex features. Results confirm that transfer learning with models like BERT, BioBERT, RoBERTa, and T5 is effective for general NLP tasks and domain-specific and multilingual sentiment analysis [20,21]. Their ability to generalize across various tasks stems from pre-training on large-scale

corpora, while fine-tuning enables adaptation to specific domains and languages. This approach yields strong results, especially in resource settings. Ongoing progress in transfer learning and techniques such as attention mechanisms and hybrid architectures is expected to further improve sentiment classification across different domains.

Table I offers a comprehensive overview of the latest sentiment analysis techniques by comparing four main categories: traditional machine learning, rule and lexicon-based approaches, deep learning models, and transfer learning methods. This summary highlights key trade-offs regarding accuracy, computational efficiency, explainability, data needs, and generalization capabilities.

Multilingual sentiment analysis is a rapidly developing area in NLP that aims to identify sentiments in texts across various languages and cultural contexts. As globalization accelerates and digital communication crosses national and language borders, the ability to gather sentiment insights from diverse sources has become essential. Multilingual skills enhance machine learning models' versatility, expanding their usefulness beyond English. Including multiple languages in sentiment analysis research addresses the need for inclusive models that are representative of global linguistic and sentiment diversity.

Multilingual sentiment analysis is challenging due to structural differences among languages. Each language has its own morphological and syntactic rules, making it insufficient to translate content into one language for analysis [22]. For example, sentiment models designed for English may not effectively capture the nuances of Arabic or Turkish without language-specific training and data enhancement. Moreover, different languages require specialized tools, such as unique part-of-speech taggers and lemmatizers, to accurately interpret negations and idiomatic expressions [3,22]. Studies indicate that building effective sentiment analysis systems for multiple languages requires either training separate models or significantly adapting a single model to match each language's unique features [23,3]. These efforts are resource-intensive and require careful consideration of idiomatic phrases, cultural contexts, and domain-specific terminology.

Recent studies highlight the importance of developing sentiment analysis models that operate effectively across multiple languages. In [24], researchers utilized multilingual BERT for sentiment classification on the Kurdish Medical Corpus, demonstrating that transformer-based architectures can help address resource gaps for underrepresented languages. Similarly, [25] tackled sentiment analysis challenges in Marathi with XLM-RoBERTa, showcasing the potential of large multilingual transformer models to perform well in low-resource settings. [26] Introduced a hybrid model combining Naive Bayes and Logistic Regression (NB-LR) for sentiment analysis in Serbian, illustrating the benefits of hybrid approaches for low-resource languages. In [27], multilingual sentiment analysis for English, Hindi, and Urdu used weak supervision with BERT, proving the effectiveness of weak supervision and the robustness of transformer models. Authors of [28] employed BiLSTM with Self-Attention and Joint Dual Input Learning (JDIL) for sentiment analysis in Hindi and Bengali, revealing the usefulness of BiLSTM architectures despite lower accuracy, which points to challenges in handling linguistic diversity. Finally, [29] extended transfer learning to Urdu sentiment analysis, emphasizing manual annotation and inter-annotator reliability as key factors for building reliable datasets in low-resource languages.

3 Research Methodology

This study aims to evaluate how advanced ML methods can enhance the performance of multilingual sentiment analysis by optimizing a pre-trained BERT model's performance. The methodology covers a review of the datasets used, a detailed analysis of feature extraction techniques, and an overview of the machine learning and deep learning models examined.

3.1 Studied Datasets

All of the studied datasets are publicly available on the Kaggle dataset website, including: "330K Arabic Sentiment Reviews Dataset" [31], "Arabic Company Reviews" [32], "Brazilian Portuguese Sentiment Analysis Datasets" [33], "Duygu Analizi Veri Seti" [34], and "Sentiment Analysis for Mental Health" [35]. The specifications of the five datasets are presented in Table 2.

Table 1. Comparative summary of sentiment analysis approaches [8]

Aspect	Traditional ML	Rule / Lexicon-based	Deep learning	Transfer learning
Accuracy	Moderate; depends on the dataset and feature engineering	Moderate in domain-specific tasks	High with sufficient labeled data	Very high across domains and tasks
Computational cost	Low	Very low	High, especially during training	Very high due to pre-training and fine-tuning
Explainability	Moderate	High (transparent, rule-based logic)	Low; black-box typically	Low; often difficult to interpret
Data requirements	Requires labeled training data	No training data; depends on lexicon quality	Requires large labeled datasets	Effective with a few labeled examples
Domain adaptability	Limited without retraining	Weak due to static lexicons	Moderate with domain-specific tuning	Strong, adaptable through fine-tuning
Robustness	Sensitive to noise and domain shift	Poor for dynamic or informal language	Improved with hybrid or regularized models	High; handles multilingual and domain-shifted data well

Table 2. The specifications of the five studied datasets

Dataset Name	330K Arabic Sentiment Reviews Dataset	Arabic Company Reviews	Brazilian Portuguese Sentiment Analysis Datasets	Duygu Analizi Veri Seti	Sentiment Analysis for Mental Health
Language(s)	Mix of Arabic and English	Arabic	Brazilian Portuguese	Turkish	English
Number of reviews	330,000	36,480	1,048,085	11,426	52,645
Number of Sentiments	2	3	5	3	7
Sentiment Categories	1 (Like) 0 (Dislike)	Positive Neutral Negative	1-star 2-star 3-star 4-star 5-star	Olumlu (Positive) Tarafsız (Neutral) Olumsuz (Negative)	Depression Stress Suicidal Anxiety Bipolar Normal Personality Disorder
Category distribution	166,853 163,147	21,214 1,819 13,447	177,834 50,028 78,548 136,695 605,470	4,191 2,867 4,371	15,385 2,667 10,643 3,888 2,876 16,348 1,200
number of words in the dataset	21,352,149	368,216	14,807,043	248,188	5,899,121

3.2 Advanced Feature Extraction Techniques

The feature extraction phase of our methodology uses both ML and DL methods to capture different aspects of sentiment-rich text. The following techniques are employed on BERT features:

3.2.1 Correlation

Correlation measures the strength and direction of relationships between signals.

- Spectral Correlation [36] provides more comprehensive features than the ordinary power spectral density. Cyclostationary analysis uncovers the hidden periodicity in the signal and offers a second-order statistical description in the frequency domain [37]. defined by the following formula:

$$S_{x_T}^{\alpha}(n,f)_{\Delta T} = \frac{1}{T}\left\langle X_T\left(n, f+\frac{\alpha}{2}\right), X_T^*\left(n, f-\frac{\alpha}{2}\right)\right\rangle_{\Delta T}$$

the correlation of the spectral components of $x(n)$ over a time span of Δt seconds,

- Cross-correlation helps determine the similarity between two signals as a function of time-lag applied to one of them. The following formula defines it:

$$R_{xy}(\tau) = \sum_n x[n]\ y[n+\tau]$$

the cross-correlation function between signals x and y at a time period m.

Correlation analysis is used for signal detection, time-delay estimation, and quality assessment in signal processing. Correlation measures how similar two signals are, which helps determine how much one signal is related to another. This is useful in time-series analysis or matching signals across different fields.

3.2.2 Principal Component Analysis (PCA)

PCA reduces the dimensionality of data while retaining most variance, transforming the original variables into a new set of uncorrelated variables. Mathematically represented by:

$$Y = X'\,V_k$$

where X' is the standardized X and V_k is a matrix, whose columns are the eigenvectors of the covariance matrix of X, sorted by corresponding eigenvalues in descending order.

PCA is useful for feature extraction, noise reduction, and visualization in high-dimensional datasets. Additionally, its simplification speeds up processing times and lowers computational costs, making data easier to visualize and analyze.

3.2.3 Linear Discriminant Analysis (LDA)

LDA classifies data by finding a linear combination of features that minimizes intra-class variance and maximizes inter-class variance, summarized as a generalized eigenproblem:

$$S_B w = \lambda S_W w \tag{4}$$

Where, S_B is Between-Class Scatter Matrix:

$$S_B = \sum_i N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

and S_W Within-Class Scatter Matrix:

$$S_W = \sum_i \sum_{x \in Ci} x(x - \mu_i)(x - \mu_i)^T \tag{6}$$

μ_i is the Mean vector, and the goal is to find the projection vector W that maximizes the $J(w)$ ratio, defined as:

$$J(w) = \frac{w^T S_B\, w}{w^T S_W\, w} \tag{7}$$

LDA for benefits as Dimensionality Reduction, Improved Model Interpretability, De-correlation of Features, Feature Selection, … is commonly used in pattern recognition and classification tasks.

3.2.4 Probabilistic Principal Component Analysis (PPCA)

Probabilistic Principal Component Analysis (PPCA) is a statistical technique that extends traditional PCA by incorporating a probabilistic model. The PPCA model can be assumed as:

$$X = WZ + \mu + \epsilon,\ \ where\ \epsilon \sim N(0, \sigma^2 I)$$

Where X is the observed data matrix, Z is an $N \times$ d matrix of latent variables with d (the number of principal components) being fewer than the original dimensions, μ is the mean vector, W is the weight matrix (or loadings) and ϵ is Gaussian noise with zero mean and variance σ^2.

PPCA provides benefits such as handling missing data, robustness to noise, improved dimensionality reduction, real-time applications, and integration with other techniques. It has various applications, especially in scenarios that require more robust analysis of high-dimensional data, particularly in noisy environments.

3.2.5 Root Mean Square (RMS)

RMS is the square root of the average of the squares of a set of values, providing a measure of the magnitude of a varying signal. It provides valuable information about the energy content, average power, and overall behavior of signals, using the following formula:

$$RMS = \sqrt{1/N \sum_{i=1}^{N} x_i^2}$$

The RMS is a measure in signal processing that provides important insights into a signal's power, quality, and energy. Its capacity to precisely reflect changes in signal magnitudes makes it useful in numerous applications.

3.2.6 Mean Square Abscissa (MSA)

MSA quantifies the average squared distance of a signal from a reference point, typically the origin or the mean of the signal, using the following formula

$$MSA = \frac{1}{N} \sum_{n} x[n]^2$$

The concept is useful in analyzing signals for their power, energy, or spread.

3.2.7 Equivalent Width

Equivalent width is a measure that quantifies the width of a spectral feature relative to its height and can provide valuable insights into various characteristics of the signal, according to the following formula:

$$W = \sum_{i} \left(1 - \frac{F_i}{F_c}\right)$$

Where F_c is the continuum level and F is the flux.

It is commonly used in signal processing and spectral analysis because it offers benefits like separating signal components, providing flexibility across domains, and facilitating signal comparison.

3.2.8 Chi-Square Goodness-of-Fit

The Chi-Square Goodness-of-Fit test is a statistical method used to determine how well a given set of observed data matches an expected distribution, using the following formula

$$\chi^2 = \sum_{i} \frac{(O_i - E_i)^2}{E_i}$$

Where χ^2 is the Chi-Square statistic, O_i is the observed frequency and E_i is the expected frequency.

It can be especially useful in various signal processing applications where it is important to analyze the distribution of signal characteristics or performance metrics. It helps validate models and assess the suitability of approximations in signal processing. We evaluated how well a normal distribution fits across feature vectors.

3.2.9 Normalization

Normalization is adjusting the amplitude or scale of a signal to a standard or specified range, typically [0,1]. It is a crucial step in many signal processing tasks and applications, enabling uniformity and improving the performance of various algorithms. By the following formula:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

3.2.10 Mean Absolute Deviation (MAD)

MAD is a statistical metric that measures the spread of a dataset by averaging the absolute deviations from the mean. In signal processing, MAD provides important insights into signal variability, stability, and noise properties, as outlined by its formula

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|$$

It is generally used for evaluating the dispersion of a signal, providing robust insights into signal variations.

3.2.11 Generalized Hurst Exponent (GHE)

GHE is a statistical measure used to characterize the long-range dependence and self-similarity of time series data. The GHE is especially useful for analyzing complex signals that display fractal properties or long-range correlations in signal processing. The Generalized Hurst Exponent can be estimated using the following formula related to rescaled range analysis:

$$R(S,n) = \frac{max\big(X(t)\big) - min\big(X(t)\big)}{S}$$

Where $R(S,n)$ is the rescaled range for a time series of data points, S is the standard deviation of the data segment, and n is the number of data points in the segment. The Hurst exponent H, is estimated by:

$$H = \lim_{n \to \infty} \frac{log\ R(n)}{log\ n}$$

Benefits include long-range dependence analysis, self-similarity characterization, and understanding

signal complexity. it is widely used in finance, geophysics, and system analysis to understand the persistence or anti-persistence of fluctuations in signals.

The above mentioned feature extraction methods collectively improve text representation by capturing both the temporal and statistical features of sentiment indicators. They are especially useful in multilingual settings where language-specific nuances require strong features for effective sentiment classification.

3.3 Decision Tree Model

Various algorithms and methodologies have emerged, each focusing on optimizing different aspects of the decision-preserving mechanisms. Decision tree models have long been at the forefront of machine learning due to their interpretability, simplicity, visual clarity, and effectiveness in solving complex classification problems in various domains, including remote sensing, medical diagnostics, and text classification [45].

Decision trees are non-parametric, interpretable models that recursively partition the input space to produce a hierarchy of decision rules. The basic structure of a decision tree includes a root node, multiple decision nodes, and terminal leaf nodes. Each decision node corresponds to a test on an attribute, and leaf nodes specify outcomes such as class labels or regression values. It has various fundamental algorithms such as ID3, C4.5, C5.0, and CART (Classification and Regression Trees). We used the CART algorithm, designed for classification and regression tasks, making it a versatile tool for various applications.

The homogeneity of continuous target variables is computed as:

$$IV(D) = E(Y^2) - \big(E(Y)\big)2 = \frac{1}{n}\sum_{i=1}^{n} {y_i}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$$

An example of the visualization of the Decision Tree model for the “Brazilian Portuguese Sentiment Analysis” dataset is presented in Figures 1 and 2, for the total and 15 splits, respectively.

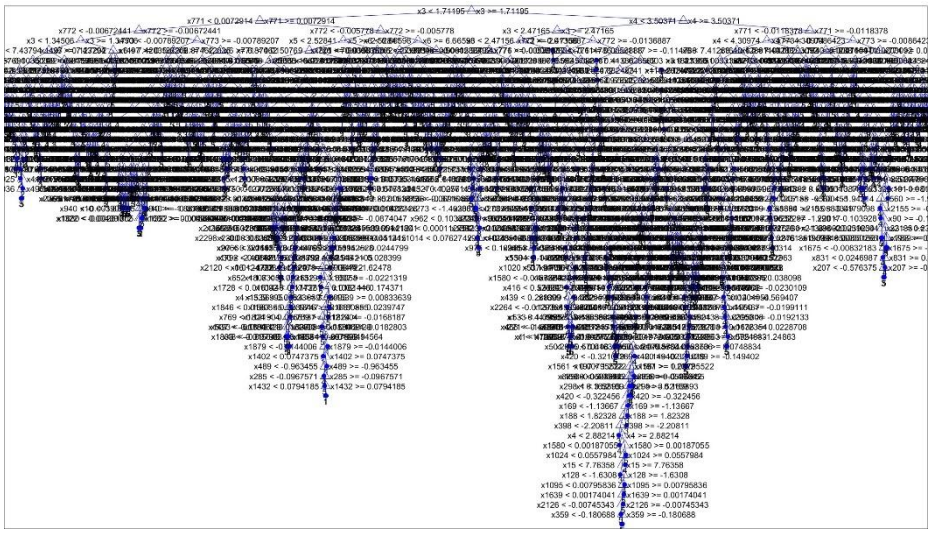


Fig. 1. Visualization of the Decision Tree model for the “Brazilian Portuguese Sentiment Analysis” dataset. (total splits)

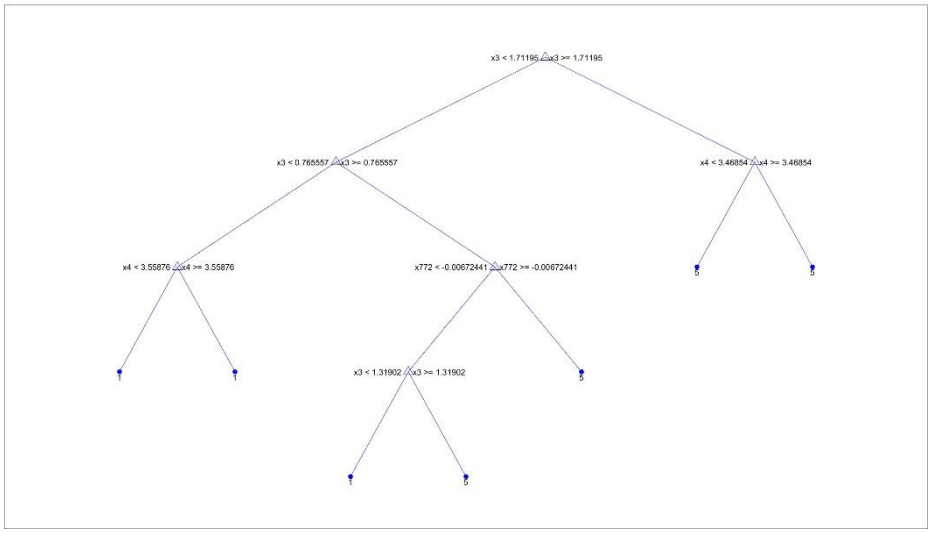


Fig. 2. Visualization of the Decision Tree model for the "Brazilian Portuguese Sentiment Analysis" dataset. (15 splits)

3.4 Bert Model

BERT (Bidirectional Encoder Representations from Transformers) is a powerful model created by Google for natural language processing (NLP). It has revolutionized NLP by introducing a bidirectional transformer mechanism that jointly conditions on both left and right contexts. With a carefully engineered pre-training phase combining masked language modeling and next sentence prediction, BERT captures syntactic, semantic, and relational knowledge. The model's internal representations dynamically transition from surface-level phrase features in lower layers to deeper semantic constructs in higher layers, proving its versatility in understanding language.

At its core, BERT integrates two complementary pre-training objectives: the Masked Language Model (MLM) objective, wherein random tokens are masked and subsequently predicted based on context, and the Next Sentence Prediction (NSP) objective, which evaluates the relationship between consecutive sentences [46]. By leveraging these strategies, BERT has enhanced downstream task performance and provided researchers with unprecedented insight into language's structural properties and relational knowledge.

4 Proposed Scheme

This section presents the proposed scheme is procedure, including: 1) text preprocessing steps, 2) text tokenizing step, 3) pre-trained BERT model, 4) advanced extracted and preprocessed features, and classification. The flowchart of the proposed method is shown in Figure 3.

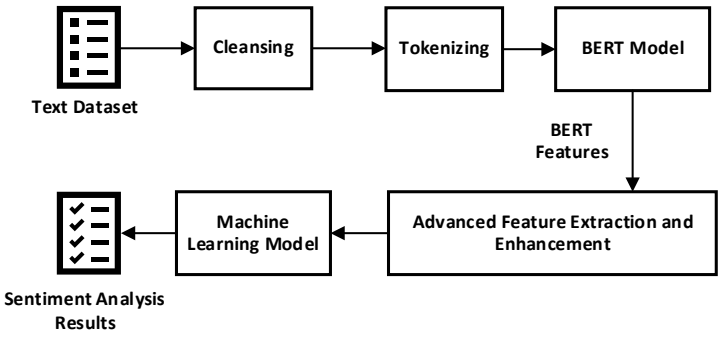


Fig. 3. The flowchart of the proposed method

4.1 Text Cleansing

Text cleansing is an important step in NLP studies. We implemented a standard cleansing pipeline, which includes:

- **Part-of-Speech (POS) tagging** is an essential process in natural language processing (NLP) that involves assigning parts of speech to each word in a sentence, such as nouns, verbs, adjectives, adverbs, and so on [38]. Recognizing the part of speech for each word can improve sentiment analysis.
- **Removing stop words** like "a", "and", "to", and "the" (known as stop words) is not useful and is considered noise in NLP. We removed these words from non-English datasets using the list of stop words provided in [39].
- **Lemmatization** is a process in natural language processing (NLP) that reduces words to their base or root form, known as the lemma [40]. For instance, in sentiment analysis, different forms of a word like "exceeded," "exceeds," and "exceeding" can be mapped to "exceed," allowing for consistent sentiment scoring.
- Unlike stemming, which might truncate words based on simple rules, lemmatization considers the context and morphological analysis of words to ensure the root form corresponds to a valid word in the language. The benefits of lemmatization include improved language understanding, increased accuracy, and reduced dimensionality.
- **Erasing punctuation** refers to the process of removing punctuation marks (including “.”, “,”, “/”, “?”, ...) from text data [41]. By removing punctuation, the text becomes more uniform, straightforward, and clearer, and reduces noise, improving the model's performance.
- **Remove words** with two or fewer characters and words with 15 or more characters [42]. Words with two or fewer characters (e.g., "a," "I," "to," "is") usually add little semantic value to the analysis. Removing them reduces noise in the dataset and helps models focus on more informative words. Words with 15 or more characters can be complex or infrequently used (e.g., "characteristically," "uncharacteristically"), which may not be relevant for all NLP tasks. These long words can complicate feature representation and lead to sparsity in the data, potentially hurting model performance. This method can boost the performance of various NLP tasks by concentrating on more meaningful and relevant words.

4.2 Tokenizing

Tokenization is a key process in NLP. It involves dividing raw text into units like words, subwords, or characters. We used the BERT tokenizer, which is based on the WordPiece algorithm—a subword method designed to split rare words into more common parts while leaving frequent words unchanged. This method helps the tokenizer cover a large vocabulary with a small set of tokens.

4.3 Pre-Trained Bert Model

BERT is a pre-trained language model trained on two major text datasets (BooksCorpus and English Wikipedia) to develop its understanding of natural language. The pre-training occurs on large text collections using self-supervised objectives before the model is fine-tuned for specific NLP tasks. This setup provides high performance across various NLP applications while minimizing data needs. We simply used a pre-trained model and did not fine-tune it on our studied datasets.

4.4 Advanced Feature Extraction and Enhancing

We extracted advanced features from BERT features and compiled them into packs, which we then evaluated using ML models. The advanced feature packs include:

- Feature pack1: Correlation (including spectral correlation and cross correlation),
- Feature pack2: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Probabilistic Principal Component Analysis (PPCA),
- Feature pack3: Other features include root mean square, mean square abscissa, equivalent width, chi-square goodness-of-fit, normalization, mean absolute deviation, and generalized Hurst exponent.

Enhancing the extracted features was achieved through combinations of PCA and LDA.

4.5 Classifying

Classifying text data using BERT and advanced BERT features was performed with the Decision Tree model. However, to improve the performance further, we used 10-fold cross-validation. K-fold cross-validation is a technique that divides the data into k subsets to provide reliable estimates of model performance. In other words, k-fold cross-validation helps ensure that the model is tested on diverse data subsets.

5 Results And Discussions

In this section, the simulation results and related discussions are presented.

5.1 RESULTS

Text cleansing, as the first step in NLP processing, improves quality by removing elements that do not contribute to the desired outcomes, which reduces the bag of words. Table 3 displays the percentage reduction for all five datasets. Figures 4 to 8 show the bag of words for the five datasets, before and after text cleansing, respectively.

Table 3. Wordbag reduction percent after the cleaning phase

Dataset	Wordbag Reduction
330K Arabic Sentiment Reviews	68.31%
Arabic Company Reviews	76.31%
Brazilian Portuguese Sentiment Analysis	72.10%
Duygu Analizi Veri Seti	78.20%
Sentiment Analysis for Mental Health	78.36%

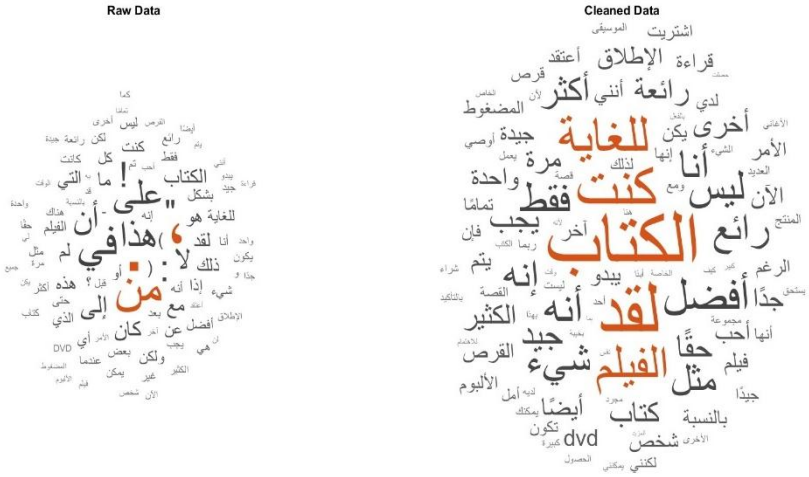


Fig. 4. The word bag of the "330K Arabic Sentiment Reviews" dataset, before and after text cleansing

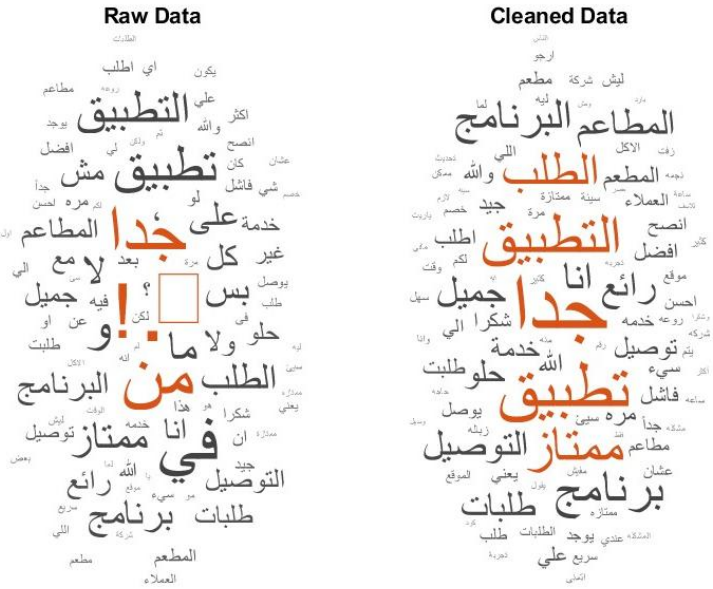


Fig. 5. The word bag of the "Arabic Company Reviews" dataset, before and after text cleansing (the non-printed word in the raw wordbag is dislike emoji (👎))

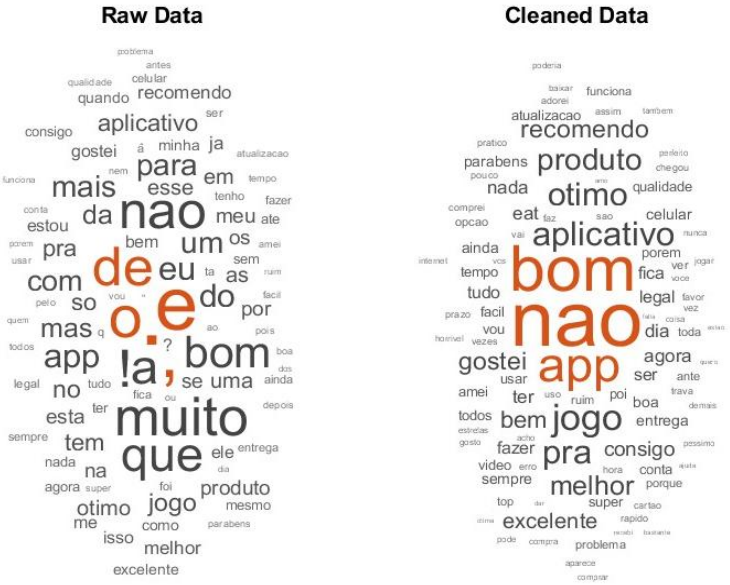


Fig. 6. The word bag of the "Brazilian Portuguese Sentiment Analysis" dataset, before and after text cleansing



Fig. 7. The word bag of the "Duygu Analizi Veri Seti" dataset, before and after text cleansing

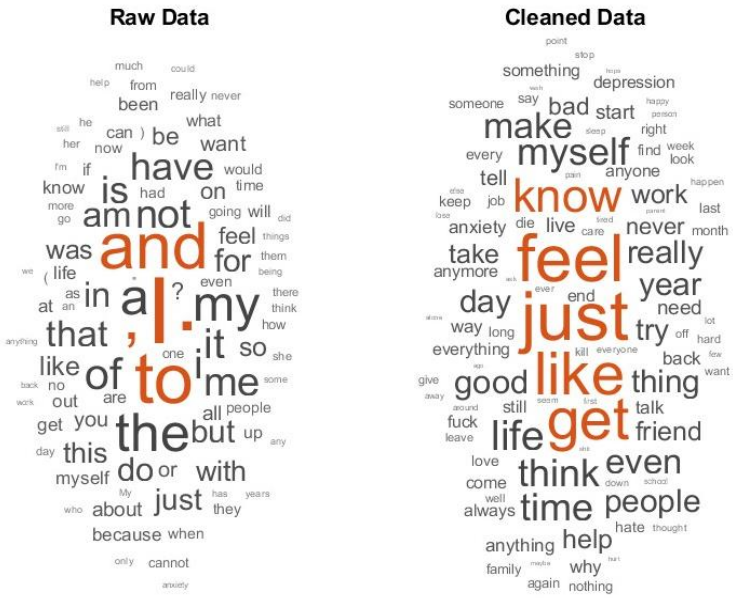


Fig. 8. The word bag of the "Sentiment Analysis for Mental Health" dataset, before and after text cleansing

Our study uses multiple evaluation metrics. For classification tasks, performance is usually measured by five key criteria:

- TP (True Positive): The model correctly predicts the positive class.
- FP (False Positive): The model predicts positive, but the true label is negative.
- TN (True Negative): The model correctly predicts the negative class.
- FN (False Negative): The model predicts negative, but the true label is positive.

So, the performance was evaluated by:

- Accuracy: Proportion of correct predictions (both positives and negatives) over all samples:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}$$

High accuracy means a large proportion of predictions are correct.

- Precision: proportion of predicted positives that are actually positive:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

- Sensitivity (Recall): The accuracy of positive predictions:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

- F1-Score: Harmonic mean of precision and recall. Balances the trade-off between false positives and false negatives:

$$\text{F1} = 2 * \frac{(\text{Precision} * \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})}$$

A high F1-score means a strong overall performance in identifying positives with few false positives and few missed positives.

- Area Under the Curve (AUC) refers to the area under the Receiver Operating Characteristic (ROC)performance curve. It summarizes the model's ability to discriminate between the positive and negative classes across all possible classification thresholds.

The results of simulations are presented in Tables 5 to 9.

Table 5. Simulation results for the Decision Tree model, "330K Arabic Sentiment Reviews" dataset

			Acc.	F1Score	AUC	Prec.	Sens.
BERT features	Train	not preprocessed	53%	52%	53%	53%	53%
		LDA+PCA preprocessed	60%	60%	60%	60%	60%
		PCA+LDA preprocessed	62%	62%	63%	62%	62%
	Valid.	not preprocessed	52%	52%	52%	52%	52%
		LDA+PCA preprocessed	66%	66%	66%	66%	66%
		PCA+LDA preprocessed	**69%**	**68%**	**68%**	**69%**	**69%**
Feature Pack 1	Train	not preprocessed	52%	52%	52%	52%	52%
		LDA+PCA preprocessed	58%	58%	59%	58%	58%
		PCA+LDA preprocessed	63%	62%	61%	63%	62%
	Valid.	not preprocessed	51%	50%	51%	51%	50%
		LDA+PCA preprocessed	53%	53%	54%	53%	53%
		PCA+LDA preprocessed	**85%**	**84%**	**88%**	**85%**	**85%**
Feature Pack 2	Train	not preprocessed	70%	43%	57%	55%	60%
		LDA+PCA preprocessed	67%	41%	55%	51%	56%
		PCA+LDA preprocessed	76%	52%	67%	63%	69%
	Valid.	not preprocessed	70%	42%	57%	55%	59%
		LDA+PCA preprocessed	66%	38%	54%	50%	54%
		PCA+LDA preprocessed	**81%**	**63%**	**77%**	**71%**	**76%**
Feature Pack 3	Train	not preprocessed	61%	61%	60%	61%	61%
		LDA+PCA preprocessed	50%	50%	50%	50%	50%
		PCA+LDA preprocessed	68%	67%	68%	68%	68%
	Valid.	not preprocessed	69%	69%	68%	69%	69%
		LDA+PCA preprocessed	51%	51%	52%	51%	51%
		PCA+LDA preprocessed	**68%**	**67%**	**68%**	**68%**	**68%**

Table 6. Simulation results for the Decision Tree model, "Arabic Company Reviews" dataset

			Acc.	F1Score	AUC	Prec.	Sens.
BERT features	Train	not preprocessed	70%	43%	57%	55%	60%
		LDA+PCA preprocessed	75%	49%	63%	62%	68%
		PCA+LDA preprocessed	76%	51%	66%	63%	69%
	Valid.	not preprocessed	69%	41%	56%	53%	58%
		LDA+PCA preprocessed	77%	53%	68%	65%	70%

		PCA+LDA preprocessed	**78%**	**56%**	**71%**	**67%**	**72%**
Feature Pack 1	Train	not preprocessed	68%	42%	57%	53%	59%
		LDA+PCA preprocessed	66%	38%	53%	50%	54%
		PCA+LDA preprocessed	75%	51%	64%	62%	68%
	Valid.	not preprocessed	68%	41%	57%	53%	58%
		LDA+PCA preprocessed	69%	40%	55%	53%	57%
		PCA+LDA preprocessed	**80%**	**61%**	**74%**	**69%**	**74%**
Feature Pack 2	Train	not preprocessed	75%	50%	62%	62%	68%
		LDA+PCA preprocessed	64%	36%	51%	47%	52%
		PCA+LDA preprocessed	76%	53%	66%	63%	69%
	Valid.	not preprocessed	77%	55%	69%	65%	70%
		LDA+PCA preprocessed	64%	35%	50%	46%	50%
		PCA+LDA preprocessed	**84%**	**69%**	**80%**	**74%**	**79%**
Feature Pack 3	Train	not preprocessed	70%	43%	57%	55%	60%
		LDA+PCA preprocessed	67%	41%	55%	51%	56%
		PCA+LDA preprocessed	76%	52%	67%	63%	69%
	Valid.	not preprocessed	70%	42%	57%	55%	59%
		LDA+PCA preprocessed	66%	38%	54%	50%	54%
		PCA+LDA preprocessed	**81%**	**63%**	**77%**	**71%**	**76%**

Table 7. Simulation results for the Decision Tree model, "Brazilian Portuguese Sentiment Analysis" dataset

			Acc.	F1Score	AUC	Prec.	Sens.
BERT features	Train	not preprocessed	73%	29%	55%	37%	54%
		LDA+PCA preprocessed	75%	32%	58%	41%	58%
		PCA+LDA preprocessed	75%	34%	60%	42%	59%
	Valid.	not preprocessed	73%	28%	56%	37%	53%
		LDA+PCA preprocessed	78%	37%	63%	46%	62%
		PCA+LDA preprocessed	**80%**	**43%**	**68%**	**49%**	**66%**
Feature Pack 1	Train	not preprocessed	71%	27%	54%	36%	52%
		LDA+PCA preprocessed	75%	31%	56%	41%	57%
		PCA+LDA preprocessed	77%	36%	60%	45%	60%
	Valid.	not preprocessed	71%	26%	53%	35%	51%
		LDA+PCA preprocessed	78%	39%	65%	47%	61%
		PCA+LDA preprocessed	**88%**	**65%**	**84%**	**68%**	**80%**
Feature Pack 2	Train	not preprocessed	75%	33%	57%	42%	57%
		LDA+PCA preprocessed	71%	24%	50%	34%	47%
		PCA+LDA preprocessed	78%	40%	65%	47%	63%
	Valid.	not preprocessed	79%	41%	66%	48%	64%
		LDA+PCA preprocessed	71%	24%	53%	34%	47%
		PCA+LDA preprocessed	**93%**	**78%**	**93%**	**78%**	**89%**
Feature Pack 3	Train	not preprocessed	74%	23%	56%	26%	46%
		LDA+PCA preprocessed	72%	21%	54%	24%	42%
		PCA+LDA preprocessed	74%	25%	58%	27%	48%
	Valid.	not preprocessed	75%	27%	61%	29%	50%
		LDA+PCA preprocessed	74%	25%	58%	26%	44%
		PCA+LDA preprocessed	**82%**	**41%**	**73%**	**41%**	**63%**

Table 8. Simulation results for the Decision Tree model, "Duygu Analizi Veri Seti" dataset

			Acc.	F1Score	AUC	Prec.	Sens.

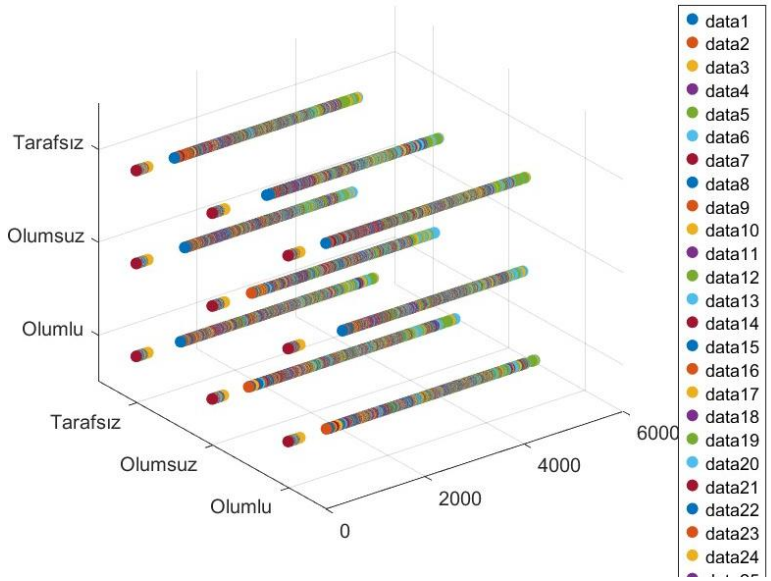
BERT features	Train	not preprocessed	60%	42%	56%	41%	45%
		LDA+PCA preprocessed	61%	42%	57%	42%	46%
		PCA+LDA preprocessed	69%	55%	67%	54%	59%
	Valid.	not preprocessed	58%	38%	54%	38%	42%
		LDA+PCA preprocessed	66%	50%	64%	50%	54%
		PCA+LDA preprocessed	**79%**	**69%**	**79%**	**67%**	**74%**
Feature Pack 1	Train	not preprocessed	58%	39%	55%	39%	43%
		LDA+PCA preprocessed	61%	43%	58%	42%	46%
		PCA+LDA preprocessed	71%	57%	68%	56%	61%
	Valid.	not preprocessed	59%	40%	55%	40%	44%
		LDA+PCA preprocessed	59%	40%	55%	39%	43%
		PCA+LDA preprocessed	**90%**	**86%**	**95%**	**83%**	**90%**
Feature Pack 2	Train	not preprocessed	69%	53%	66%	53%	57%
		LDA+PCA preprocessed	56%	35%	50%	35%	37%
		PCA+LDA preprocessed	75%	63%	74%	62%	66%
	Valid.	not preprocessed	**78%**	**66%**	**76%**	**65%**	**70%**
		LDA+PCA preprocessed	55%	34%	50%	34%	37%
		PCA+LDA preprocessed	75%	63%	74%	62%	66%
Feature Pack 3	Train	not preprocessed	60%	41%	56%	40%	44%
		LDA+PCA preprocessed	71%	56%	68%	56%	60%
		PCA+LDA preprocessed	72%	58%	70%	57%	61%
	Valid.	not preprocessed	58%	38%	54%	38%	42%
		LDA+PCA preprocessed	56%	35%	50%	35%	38%
		PCA+LDA preprocessed	**88%**	**83%**	**92%**	**79%**	**87%**

Table 9. Simulation results for the Decision Tree model, "Sentiment Analysis for Mental Health" dataset

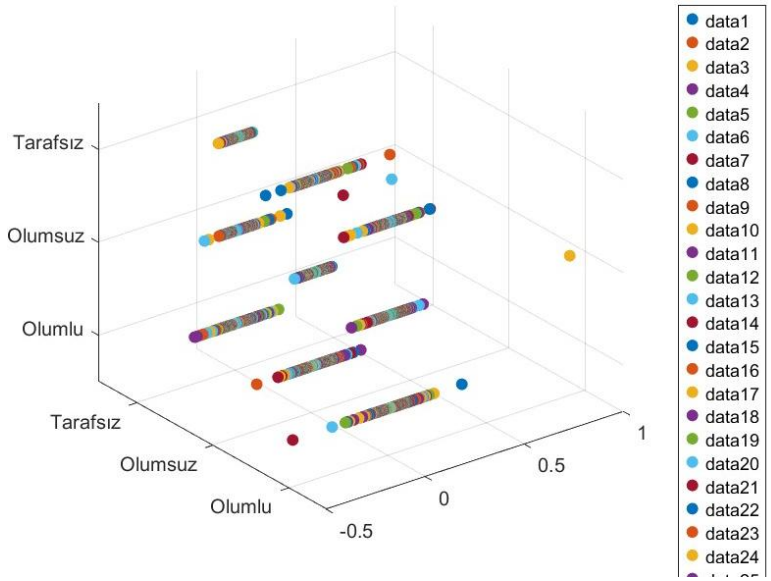
			Acc.	F1Score	AUC	Prec.	Sens.
BERT features	Train	not preprocessed	74%	24%	56%	27%	47%
		LDA+PCA preprocessed	74%	24%	57%	27%	48%
		PCA+LDA preprocessed	74%	25%	58%	28%	49%
	Valid.	not preprocessed	78%	31%	65%	34%	55%
		LDA+PCA preprocessed	83%	41%	72%	43%	63%
		PCA+LDA preprocessed	**83%**	**44%**	**75%**	**44%**	**66%**
Feature Pack 1	Train	not preprocessed	74%	24%	57%	26%	47%
		LDA+PCA preprocessed	75%	24%	56%	28%	48%
		PCA+LDA preprocessed	75%	26%	58%	29%	49%
	Valid.	not preprocessed	76%	28%	61%	30%	52%
		LDA+PCA preprocessed	83%	40%	72%	43%	62%
		PCA+LDA preprocessed	**86%**	**50%**	**79%**	**50%**	**70%**
Feature Pack 2	Train	not preprocessed	75%	25%	57%	28%	48%
		LDA+PCA preprocessed	71%	18%	51%	20%	36%
		PCA+LDA preprocessed	76%	27%	59%	30%	50%
	Valid.	not preprocessed	83%	42%	74%	44%	64%
		LDA+PCA preprocessed	72%	19%	53%	21%	37%
		PCA+LDA preprocessed	**88%**	**56%**	**82%**	**56%**	**74%**
Feature Pack 3	Train	not preprocessed	75%	24%	56%	27%	47%
		LDA+PCA preprocessed	73%	22%	55%	24%	42%
		PCA+LDA preprocessed	75%	26%	58%	29%	49%
	Valid.	not preprocessed	78%	29%	62%	32%	53%
		LDA+PCA preprocessed	82%	41%	72%	41%	61%
		PCA+LDA preprocessed	**85%**	**49%**	**78%**	**49%**	**69%**

As detailed in Tables 5 to 9, our approach significantly improves the performance of a BERT-based model, without altering the core BERT architecture. The results align with the expected performance of ML-based NLP methods shown in Table 1: moderate, and dependent on the dataset and feature engineering.

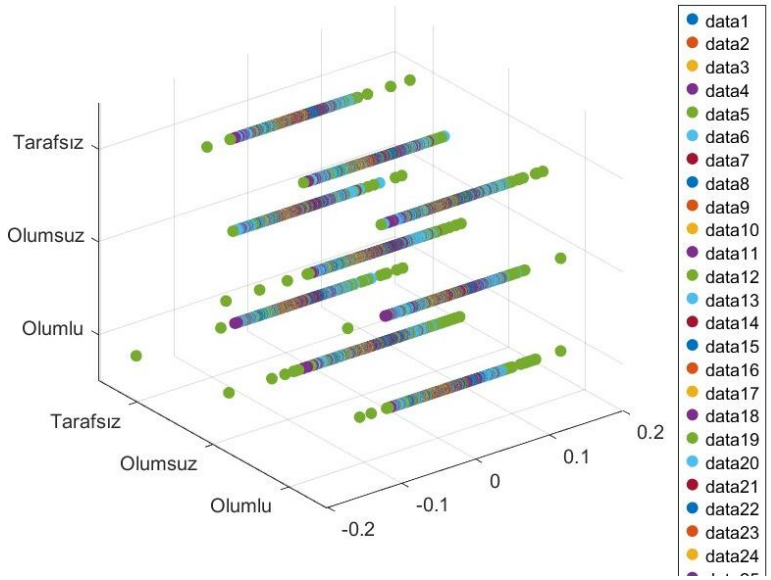
In a comprehensive overview, Feature Pack 1 (correlation features, including spectral correlation and cross-correlation features derived from BERT features) and Feature Pack 2 (PCA, LDA, and PPCA features obtained from BERT features) demonstrated the best performance compared to the others. Correlation-based feature extraction involves assessing the relationships among features obtained from BERT embeddings. Cross-correlation examines these relationships in the time domain. In contrast, spectral correlation analysis (cyclostationary analysis in this study) reveals hidden periodic patterns in the signal and provides a second-order statistical description in the frequency domain. Since BERT employs attention mechanisms to identify contextual and semantic relationships among words, quantifying the correlations among these features can yield valuable insights for sentiment detection. The scatter plot of the Correlation analysis preprocessing for the "Duygu Analizi Veri Seti" dataset is presented in Figure 9.



a. not processed features



b. PCA processed features



c. LDA processed features

Fig. 9. Scatter plot of Correlation analysis preprocessing for "Duygu Analizi Veri Seti" dataset

From another point of view, Correlation-based features enhanced the performance of the model by:

- **Capturing Interdependencies**: While BERT embeddings already encode contextual relationships, applying correlation metrics further highlights how different dimensions interact. This helps identify clusters of features that collectively contribute to sentiment classification.
- **Enhancing Discriminative Power**: By quantifying which features are highly interdependent, the model can emphasize sentiment-informative dimensions. This leads to a refined feature subset that is more predictive of sentiment classes, boosting accuracy and robustness.
- **Noise Reduction**: Correlation-based techniques can help in filtering out redundant or noisy features that do not contribute meaningfully to the task. This

pruning ensures that only the most relevant aspects of BERT embeddings are retained for final classification.

On the other hand, feature preprocessing on almost all of the feature packs had significantly improved the performance and speed of the classification model, which highlights the influence of feature enhancement on such applications.

Figure 10 illustrate the improvements in sentiment analysis across various datasets by comparing a pre-trained BERT model, with and without the proposed method. The base accuracy represented in the chart corresponds to the BERT model, while the second percentage shows the maximum gain achieved by the proposed approach for each dataset.

6 Discussions

6.1 t-SNE analysis

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a machine learning technique used for visualizing high-dimensional data by reducing it to a lower-dimensional space. Applying t-SNE analysis to different feature types in datasets reveals strong correlations among features. This highlights the importance of using sentiment analysis methods, such as TextBlob and lexicon-based techniques, to enhance the accuracy of NLP applications.

For example, t-SNE analysis results for the extracted features of the "330K Arabic Sentiment Reviews" dataset are presented in Figure 11.

6.2 Dataset issues

Dataset pre-processing is a crucial step in enhancing the method's performance. A comprehensive pre-processing pipeline ensures the dataset is prepared for analysis and model training. We implemented several key techniques as discussed in the "text cleansing" section. In future studies, we also intend to explore additional methods such as managing class imbalance and data augmentation (to increase dataset diversity and size), handling missing values (by removing incomplete or blank sentiments), and addressing data anomalies like outliers, rare words or phrases, unusual patterns, discrepancies in sentiment, and inconsistent chatbot behavior. Because of our machine learning approach, we did not correct these anomalies and proceeded with the original datasets.

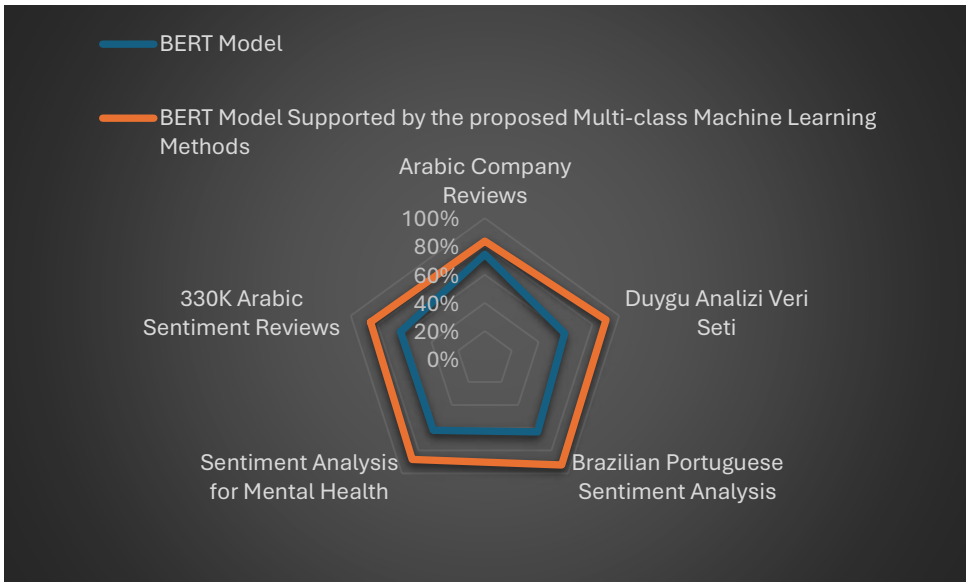
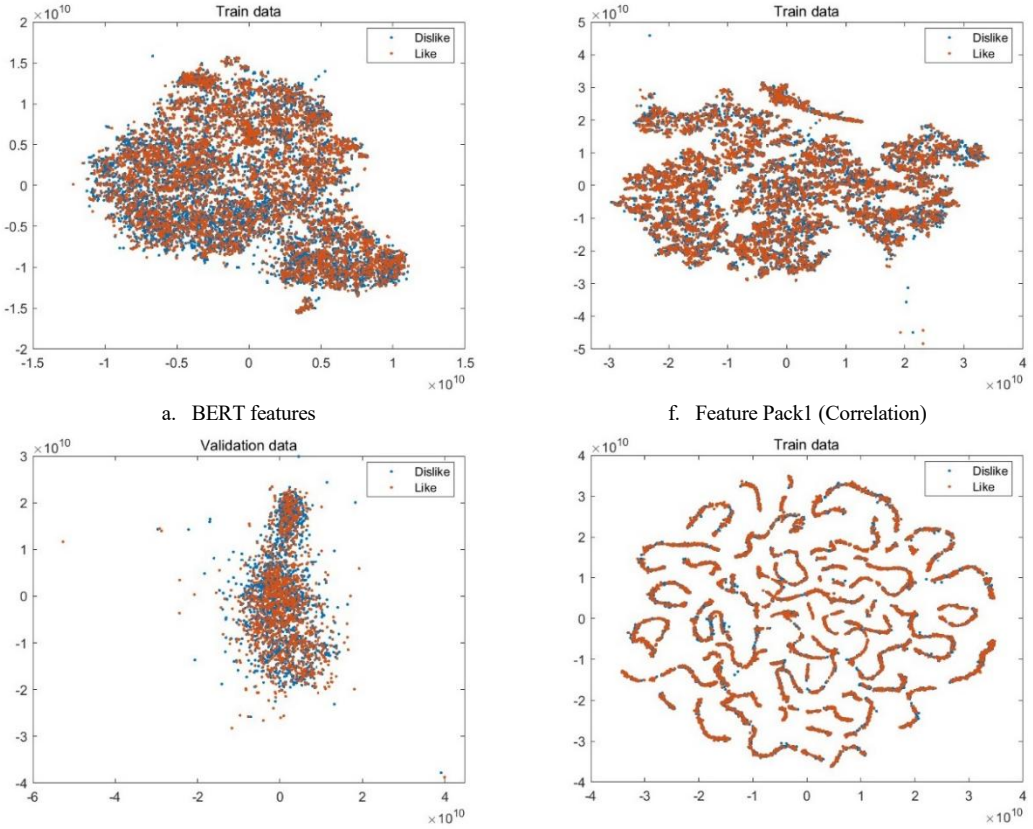


Fig. 10. The accuracy improvement achieved by the proposed method



a. BERT features

f. Feature Pack1 (Correlation)

g. Feature Pack2 (PCA, LDA, PPCA)

h. Feature Pack3 (RMS, MSA, …)

Fig. 11. t-SNE analysis for the extracted features of "330K Arabic Sentiment Reviews" Dataset

7 Conclusion And Suggestion For Feature Works

7.1 Conclusion

Recent trends in sentiment analysis have seen the rise of transformer-based models such as BERT, which leverage deep bidirectional representations to capture contextual nuances. BERT and its variants have demonstrated impressive results across multiple tasks, including sentiment analysis, due to their ability to incorporate extensive contextual information and handle large-scale data. However, these models also have inherent limitations, such as high computational cost and the necessity for large training datasets to achieve optimal performance. Simultaneously, classical models continue to offer competitive performance in scenarios where computational resources are limited or when rapid deployment is necessary. These models typically benefit from simpler feature sets derived from advanced feature extraction and preprocessing, creating a compelling case for continued inclusion in comparative studies.

We got up to 28% improvement in the accuracy of sentiment analysis, and the results demonstrated the effectiveness of Machine Learning (ML) techniques. The approach significantly improves the performance of a BERT-based model without altering the core Deep learning (DL) architecture. The performance is enhanced and aligns with the expected performance of the ML-based method, moderately and dependently on the dataset and feature engineering. This emphasizes the importance of DL and Transfer learning models in Natural Language Processing (NLP).

A pre-trained BERT model is mainstream for capturing semantic and contextual information within texts. Then, various advanced feature extraction, feature enhancement, and a 10-fold cross-validated Decision Tree model were used to conduct the proposed method. The results indicated the effectiveness of correlation analysis and dimensionality reduction techniques.

Correlation-based feature extraction involves analyzing the relationships between extracted features from the BERT model. Cross-correlation examines these relationships in the time domain. In contrast, spectral correlation analysis (cyclostationary analysis in this study) reveals hidden periodic patterns in the signal and provides a second-order statistical description in the frequency domain. Since BERT employs attention mechanisms to identify contextual and semantic relationships among words, quantifying the correlations among these features can yield valuable insights for sentiment detection.

Dimensionality Reduction Techniques such as PCA and LDA reduce computational complexity and eliminate noise and irrelevant components, enhancing model robustness and generalization.

Moreover, Multilingual sentiment analysis is challenging due to structural differences among languages. Each language has its own morphological and syntactic rules, making it insufficient to translate content into one language for analysis. Thus, a Multilingual approach that integrates innovative feature extraction methods and rigorous preprocessing is essential for practical multilingual sentiment analysis. The proposed scheme showed good performance on various languages.

7.2 Suggestion For Future Work

Our future work will focus on deep learning and Transfer learning methods. Additionally, incorporating pre-trained and fine-tuned models on the under-study dataset or domain (such as sentiment analysis, pharmaceutical data interpretation, and radiology report studies), along with effective augmentation methods and careful dataset cleaning, animalities reduction, and sentiment analysis techniques, will significantly improve the performance of the NLP model.

Declarations

- Conflict of interest: The authors declare that they have no conflict of interest.
- Data availability: the studied datasets are publicly available at: https://www.kaggle.com/
- Funding statement: there is no funding.

Refferences

1. J. Francis and M. Subha, "An Overview of Natural Language Processing (NLP) in Healthcare: Implications for English Language Teaching," 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 2024, pp. 824-827, doi: 10.1109/I-SMAC61858.2024.10714890.
2. Shiv Singh, "Leading Your Business in the AI Era," in Marketing with AI For Dummies , Wiley, 2025, pp.319-333
3. Basant, Agarwal., Namita, Mittal, Machine Learning Approaches for Sentiment Analysis. 1740-1756. doi: 10.4018/978-1-4666-6086-1.CH011, 2014
4. V. Ganganwar and R. Rajalakshmi, “Implicit aspect extraction for sentiment analysis: A survey of recent approaches,” Proc. Comput. Sci., vol. 165, pp. 485–491, Jan. 2019, doi: 10.1016/j.procs.2020.01.010.
5. Hasan M, Islam I, Hasan KA, Sentiment analysis using out of core learning. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp 1–6. IEEE
6. Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P (2019) Understanding emotions in text using deep learning and big data. Comput Hum Behav 93:309–317
7. Ziru Fu; Yu Cheng Hsu; Christian S Chan; Chaak Ming Lau; Joyce Liu; Paul Siu, Fai Yip, “Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study”, JOURNAL OF MEDICAL INTERNET RESEARCH, J Med Internet Res 2024 | vol. 26 | e51069 | p. 1, https://www.jmir.org/2024/1/e51069.
8. Alahmadi, K., Alharbi, S., Chen, J. et al. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. Soc. Netw. Anal. Min. 15, 45 (2025). https://doi.org/10.1007/s13278-025-01461-8
9. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. arXiv preprint cs/0205070
10. Fei Y, Simultaneous support vector selection and parameter optimization using support vector machines for sentiment classification. In: 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp 59–62. IEEE
11. Gopalakrishnan V, Ramaswamy C, Sentiment learning from imbalanced dataset: an ensemble based method. Int J Artif Intell, 2014, 12(2):75–87
12. Tripathy A, Agrawal A, Rath SK, Classification of sentiment reviews using n-gram machine learning approach. Expert Syst Appl 2016, 57:117–126
13. Srivastava A et al, Improving sentiment analysis with neural networks. In: International Conference on Knowledge-Based Organization, 2020
14. Fadel R, Öz K, A sentiment analysis model for terrorist attacks reviews on twitter using a hybrid lexicon-machine learning approach. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp 163–170
15. Kim Y, Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014
16. H. Wang, Y. Zhu, Z. Song, and T. Chen, “A Multi-Scale Local Attention Network for Image Semantic Segmentation,” IEEE Transactions on Image Processing, vol. 33,

pp. 1721–1733, Jan. 2024. doi: 10.1109/TIP.2023.3334456
17. P. Chen, X. Li, and J. Zhang, “An Enhanced LSTM Model for Improving Sequential Data Classification,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 5, pp. 2378–2389, May 2024. doi: 10.1109/TNNLS.2024.3326789.
18. Chen B, Huang Q, Chen Y, Cheng L, Chen R, Deep neural networks for multi-class sentiment classification. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp 854–859. IEEE
19. Abdullah T, Ahmet A, Deep learning in sentiment analysis: recent architectures. ACM Comput Surv 55(8):1–37, 2022
20. Zhu X, Chen Y, Gu Y, Xiao Z, Sentimedqaer: a transfer learning-based sentiment-aware model for biomedical question answering. Front Neurorobot 16:773329, 2020
21. Prottasha NJ, Sami AA, Kowsher M, Murad SA, Bairagi AK, Masud M, Baz M, Transfer learning for sentiment analysis using bert based supervised fine-tuning. Sensors 22(11):4157, 2022
22. Felix, Greaves., Daniel, Ramirez-Cano., Christopher, Millett., Ara, Darzi., Liam, Donaldson, Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. Journal of Medical Internet Research, 15(11) doi: 10.2196/JMIR.2721, 2013
23. B, Schuller., J-G, Ganascia., Laurence, Devillers, Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation., 2016
24. Badawi SS, Using multilingual bidirectional encoder representations from transformers on medical corpus for kurdish text classification. Aro-Sci J Koya Univ 11(1):10–15, 2023
25. Rathod N, Mistry N, Talati D, Parikh M, Kore A, Kanani P, Marathi social media opinion mining using xlm-r. In: 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pp 730–736. IEEE
26. Draskovic D, Zecevic D, Nikolic B, Development of a multilingual model for machine sentiment analysis in the serbian language. Mathematics 10(18):3236, 2022
27. Rastogi S, Weak supervision and transformed-based sentiment analysis on multi-lingual data. In: 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp 706–712. IEEE
28. Khan S, Shahid M, Hindi/bengali sentiment analysis using transfer learning and joint dual input learning with self attention. arXiv preprint arXiv:2202.05457, 2022
29. Ahmed AAA, Agarwal S, Kurniawan IGA, Anantadjaya SP, Krishnan C, Business boosting through sentiment analysis using artificial intelligence approach. Int J Syst Assurance Eng Manag 13(Suppl 1):699–709, 2022
30. Badawi SS, Using multilingual bidirectional encoder representations from transformers on medical corpus for kurdish text classification. Aro-Sci J Koya Univ 11(1):10–15, 2023
31. https://www.kaggle.com/datasets/abdallaellaithy/330k-arabic-sentiment-reviews
32. https://www.kaggle.com/datasets/fahdseddik/arabic-company-reviews
33. https://www.kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets
34. https://www.kaggle.com/datasets/burhanbilenn/duygu-analizi-icin-urun-yorumlari
35. https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health
36. W. A. Gardner, ‘‘The spectral correlation theory of cyclostationary timeseries,’’ Signal Process., vol. 11, no. 1, pp. 13–36, 1986.
37. N. Panahi, M. C. Amirani and M. Valizadeh, "An EEG-Based Brain-Computer Interface Using Spectral Correlation Function," in IEEE Access, vol. 11, pp. 33236-33247, 2023, doi: 10.1109/ACCESS.2023.3262465.
38. Chotirat S, Meesad P. Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning. Heliyon. 2021 Oct 19;7(10):e08216. doi: 10.1016/j.heliyon.2021.e08216. PMID: 34746470; PMCID: PMC8554172.
39. https://www.kaggle.com/datasets/heeraldedhia/stop-words-in-28-languages
40. Kumar, A., & Singh, S., "A Survey on Text Preprocessing Techniques in Natural Language Processing", Journal of King Saud University - Computer and Information Sciences, 2023, DOI: 10.1016/j.jksuci.2023.03.007
41. Chen, Y., & Wang, L."The Impact of Punctuation on Emotion Detection in Text: A Machine Learning Approach", IEEE Transactions on Affective Computing, 2024, DOI: 10.1109/TAFFC.2024.1234567
42. Kumar, A., & Sharma, S., "Text Preprocessing Techniques for Natural Language Processing: A Comprehensive Review", International Journal of Information Technology, 2024, DOI: 10.1007/s41870-024-00888-5
43. A. Albladi, M. Islam and C. Seals, "Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review," in IEEE Access, vol. 13, pp. 30444-30468, 2025, doi: 10.1109/ACCESS.2025.3541494
44. https://www.mathworks.com/help/stats/fitcecoc.html
45. Muhammad Asfand Hafeez, et al, Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm, Applied Science, 2021, https://doi.org/10.3390/app11156728.
46. Ganesh Jawahar,et al., What does BERT learn about the structure of language?, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy, July 28 - August 2, 2019. Association for Computational Linguistics



NAZILA PANAHI received the B.Sc. degree in telecommunication engineering from Islamic Azad University (Urmia branch), Urmia, Iran, in 2008. She received her M.Sc. and PhD. degree in telecommunication engineering from Urmia University, Urmia, Iran, in 2012 and 2023, respectively.
Her research interests include signal processing, pattern recognition and biometrics.
(corresponding author)



MEHDI CHEHEL AMIRANI received the B.Sc. degree in Electronic Engineering from Urmia University, Urmia, Iran, in 1993 and the M.Sc. and Ph.D. degrees in Communication Engineering from Iran University of Science and Technology (IUST), Tehran, Iran, in 1998 and 2009, respectively. Currently, he is a professor of communication at Urmia University, Urmia, Iran.

His research interests include pattern recognition, digital signal processing, and secure communication.