International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 3, June 2026, pp. 1474~1484
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i3.pp1474-1484

Integrating BERT fine-tuning and genetic algorithm for superior depression detection in social media

Abd Allah Aouragh[1], Mohamed Bahaj[1], Fouad Toufik[2]
[1]MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco
[2]Computer Sciences Laboratory, Higher School of Technology, Mohammed V University, Sale, Morocco

Article Info

Article history:

Received Aug 29, 2024
Revised Feb 21, 2026
Accepted Mar 16, 2026

Keywords:

BERT
Depression detection
Genetic algorithm
Machine learning
Natural language processing
Support vector machine
Vectorization techniques

ABSTRACT

Early detection of depression is crucial for minimizing its adverse effects on mental and physical health. Recent advancements in natural language processing facilitate the large-scale analysis of social media texts to identify depressive tendencies. Our study introduces a novel approach by integrating a genetic algorithm for hyperparameter tuning, optimizing the classification performance beyond conventional methods. We provide a comprehensive comparison of vectorization techniques, including term frequency-inverse document frequency (TF-IDF), Word2Vec, and a fine-tuned bidirectional encoder representation from transformers (BERT) model specifically adapted to our dataset. Using a dataset of 7,731 entries, we implemented standard pre-processing steps such as stop word removal and lemmatization before vectorizing the text. Five machine learning algorithms—decision tree, logistic regression, random forest, gradient boosting, and support vector machine—were evaluated, with hyperparameter tuning performed using a genetic algorithm. The highest accuracy (95.99%) and F1-score (95.91%) were achieved with the combination of fine-tuned BERT, support vector machine, and genetic algorithm optimization. This study demonstrates the advantages of integrating BERT fine-tuning with genetic optimization, outperforming traditional TF-IDF and Word2Vec approaches in depression detection.

Corresponding Author:

Abd Allah Aouragh
MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University
Settat, Morocco
Email: abdallahaouragh@gmail.com

1. INTRODUCTION

Depression is a widespread mental illness affecting millions of people worldwide, characterized by persistent feelings of sadness, loss of interest, and decreased energy [1]. According to the World Health Organization (WHO), 5% of the global population currently suffers from depression, which, in its most severe forms, can lead to suicide [2]. Around 700,000 people worldwide took their own lives, a number that is steadily increasing, underscoring the gravity of depression as a public health issue [1], [2]. The consequences of this disease include significant harm to quality of life, social relationships, and a diminished capacity to perform daily activities [2], [3]. Treatment options encompass pharmacological therapies, psychotherapy, and lifestyle interventions [3]. However, challenges persist, such as underdiagnosis, associated stigma, and limited access to care in some regions [4]. Therefore, early detection and timely intervention are crucial to mitigating the adverse effects of depression and reducing its overall impact on public health [5].

Machine learning (ML) techniques, when integrated with advancements in natural language processing (NLP), offer powerful tools for analyzing textual data and predicting medical conditions such as depression and other diseases [6], [7]. By leveraging NLP, we can examine and interpret the subtleties of language in social network posts, facilitating the early detection of depressive symptoms [8]. Unlike general sentiment analysis, which classifies text as positive, negative, or neutral, depression detection requires identifying linguistic patterns specific to clinical depressive states, such as self-referential expressions and cognitive distortions. Sentiment analysis alone may misinterpret these cues, making it necessary to develop specialized models tailored for depression detection [6], [8]. Advanced vectorization techniques, including term frequency-inverse document frequency (TF-IDF), Word2Vec, and the fine-tuned bidirectional encoder representation from transformers (BERT) model, play a crucial role in capturing the context and subtle emotions expressed in text [9]. Moreover, hyperparameter optimization through genetic algorithms [10] significantly enhances the accuracy and performance of predictive models. Another key advantage of NLP is its capacity to continuously monitor changes in an individual's mental state via real-time text data analysis, allowing for rapid response to evolving conditions and timely support [11]. By integrating these techniques, we can not only detect depression earlier but also gain deeper insights into its manifestations and triggers, potentially leading to more effective prevention and treatment strategies.

In this context, Arachchige *et al.* [12] review NLP and ML techniques for identifying depression in online support forums, analyzing 29 articles to determine the most effective and scalable feature combinations. Their study highlights challenges such as practical implementation and ethical issues, suggesting future research to refine these approaches and enhance their clinical application. Glaz *et al.* [13] highlight the growing use of ML and NLP in medicine, noting that while these models offer innovative insights and often validate existing hypotheses, they are typically limited to specific cohorts, like social media users, which may limit their general applicability. Their survey emphasizes the potential of NLP to explore otherwise inaccessible data but stresses the need for careful ethical evaluation before integrating these techniques into mental health care. Jain *et al.* [14] applied ML and NLP to predict depressive posts on Reddit. They analyzed comments and posts related to suicidal ideation using algorithms such as naive Bayes, support vector machine (SVM), logistic regression, and random forest. The study achieved an accuracy of 77.12% and an F1-score of 77% with SVM, highlighting the effectiveness of these techniques in identifying at-risk individuals on online platforms. Saifullah *et al.* [15] investigated anxiety detection by analyzing approximately 4,862 comments from YouTube. They applied data mining and machine learning algorithms, testing six classifiers. The best performance was achieved by the Random Forest model, which attained an accuracy of 84.99%, demonstrating the effectiveness of machine learning techniques in identifying emotions from online interactions.

Kour *et al.* [16] investigated predicting users' mental states by classifying depressed and non-depressed individuals from Twitter data. They used a hybrid model combining a convolutional neural network (CNN) and a bidirectional long short-term memory network (BiLSTM), achieving an accuracy of 94.28% on depression-related tweets. This model outperformed other approaches in predictive performance. Chereddy *et al.* [17] highlighted the potential of NLP for detecting depression through tweet analysis on social networks. They noted the need for more robust data to enhance accuracy and recall. Their study, employing classifiers like naive Bayes, SVM, and logistic regression to categorize tweets into positive and negative sentiments, found that SVM achieved the highest performance with 85% accuracy, 89% precision, and 63% recall and F1-score. Chen *et al.* [18] utilized NLP to detect depression in unstructured medical records. Analyzing 22,355 Mandarin records, they employed a BERT model combined with CNNs. Their study showed strong performance with general (AUC 93%) and civilian (AUC 91%) models, but the military samples underperformed (AUC 79%), with the military-specific model achieving a better AUC of 82%. The findings validate the use of deep learning techniques for depression screening while also emphasizing the importance of considering specific contexts, such as military status. Subramanian *et al.* [19] investigated the role of social media in the spread of hostile and toxic content, including hate speech and abusive language. Their study employed textual representation and encoding techniques like bi-grams, tri-grams, and FastText to capture semantic and syntactic information. Machine and deep learning methods, including CNN, BERT, and SVM, were used to classify social media posts as either indicative of depression or not. The results showed that SVM achieved an accuracy of 80% in identifying depression symptoms.

Our research builds on existing studies of depression detection from social network posts, addressing several gaps identified in the literature. While previous approaches have often relied on standard ML techniques and conventional NLP models with limited hyperparameter optimization, our study introduces a more innovative and integrated approach. We enhance model performance through advanced techniques, including fine-tuned BERT and genetic algorithm-based hyperparameter optimization, setting our work apart from earlier methods. We compared various text vectorization methods, including TF-IDF, Word2Vec, and a BERT model specifically tuned to our dataset, enhancing the capture of linguistic nuances in online forums. Additionally, our study is notable for incorporating a genetic algorithm for hyperparameter optimization, a

technique often overlooked or addressed through simpler grid search methods in other studies. This approach allows for a more thorough exploration of the hyperparameter space, resulting in notable improvements in predictive performance. Additionally, our study evaluates a variety of machine learning algorithms, including decision tree, logistic regression, random forest, gradient boosting, and support vector machine, ensuring a comprehensive and rigorous assessment of the most effective methods for classifying Reddit posts.

The following sections are organized as follows: section 2 describes the materials and methods employed in the study. Section 3 presents and analyzes the results, highlighting the impact of the applied techniques. Section 4 explains the limitations of the study. Finally, section 5 summarizes the key conclusions and offers suggestions for future research.

2. MATERIALS AND METHODS

2.1. Proposed methodology

For our study on depression detection from Reddit posts, we utilized a dataset comprising 7,731 entries. We adopted a rigorous methodology incorporating NLP and ML techniques to ensure a systematic workflow. The process commenced with the application of traditional NLP techniques, including stop word removal and lemmatization, to prepare the textual data for detailed analysis. Next, we vectorized the texts using three distinct techniques: TF-IDF, Word2Vec, and a fine-tuned BERT model specifically adapted to our dataset. This comparison allowed us to assess the effectiveness of each vectorization method in the context of depression detection. All models were evaluated using 5-fold cross-validation to ensure robust performance measurement. We then applied five machine learning algorithms: decision tree, logistic regression, random forest, gradient boosting, and support vector machine. To enhance the performance of these models, we employed a genetic algorithm for hyperparameter optimization. This structured approach aims to maximize the accuracy of depression detection through a comparative evaluation of vectorization techniques and comprehensive model optimization. The study's framework is illustrated in Figure 1.

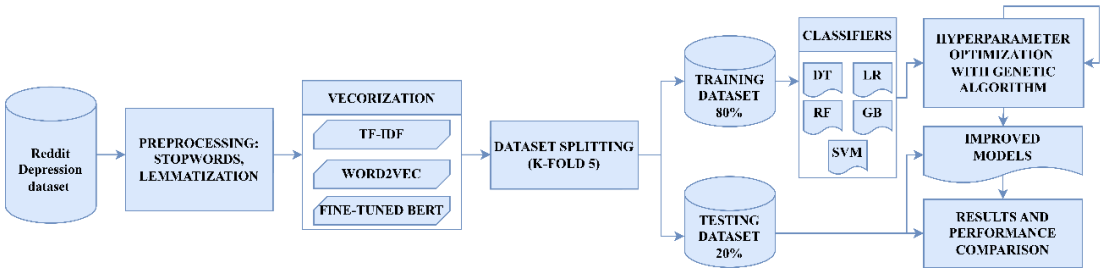

Figure 1. Overview of the proposed approach

2.2. Dataset

The dataset used in our study, available on Kaggle [20], consists of 7,731 Reddit posts, with 3,900 categorized as depressive and 3,831 as non-depressive. Each post has an average length of 368 characters. This dataset is particularly advantageous for depression detection, as Reddit posts often feature direct and spontaneous expressions of users' emotions, offering valuable insights for identifying signs of depression. The dataset is well balanced, with 50.4% of posts categorized as depressive and 49.6% as non-depressive, eliminating the need for resampling techniques to address class imbalance and contributing to a more reliable evaluation of classification models. Given this balance, no data augmentation techniques were applied, as artificial transformations, such as paraphrasing or back-translation, could introduce biases or distort the subtle linguistic markers critical for accurate classification. Since depression-related language is highly nuanced, preserving the authenticity of user-generated content ensures that the model learns from real depressive and non-depressive text samples, enhancing classification reliability. However, it is important to note that the dataset consists exclusively of English-language posts and does not provide demographic details such as age or socio-economic background, which may limit the generalizability of the findings to more diverse populations.

2.3. Preprocessing techniques

Data preprocessing is essential in natural language processing, ensuring that textual data is structured for effective analysis. In our study, we applied standard preprocessing steps, including stop word removal and lemmatization, to refine textual features and enhance classification performance [21].

2.3.1. Stop words

Stop word removal eliminates frequently occurring words (*e.g.*, “the,” “and,” “in”) that contribute little to the semantic meaning of a sentence. This step reduces data dimensionality, removes noise, and improves computational efficiency. By discarding non-informative words, the model focuses on key terms that contribute to the classification task [21].

2.3.2. Lemmatization

Lemmatization standardizes words by converting them to their base or dictionary form while preserving grammatical meaning. Unlike stemming, which trims words to their root ("running," "ran," and "runs" to "run"), lemmatization considers linguistic context, ensuring accurate word transformations. This process minimizes redundant variations in textual data, leading to more consistent and meaningful features for classification models [21].

2.4. Vectorization techniques

In NLP, vectorization is crucial for converting text into numerical data that ML algorithms can process. These methods transform words into vectors, capturing semantic and contextual relationships, and offer significant advantages in accuracy and flexibility for text classification [22], [23]. In our study, we selected the following vectorization methods for their efficiency in capturing complex patterns in large social media datasets and their widespread popularity.

2.4.1. TF-IDF

TF-IDF is a vectorization technique in NLP that assesses a word's significance within a document relative to a larger corpus. Term frequency (TF) calculates how frequently a term appears in a document, while inverse document frequency (IDF) down weights terms that are common across many documents. By balancing local frequency with global rarity, TF-IDF enhances the accuracy of textual representations, making it effective for highlighting important words in context [22], [23].

2.4.2. Word2Vec

Word2Vec is a word vectorization technique that represents words as dense vectors in a high-dimensional space. Developed by Google, this method learns these vector representations by analyzing word co-occurrences within a text corpus, ensuring that words appearing in similar contexts have closely aligned vectors. This approach effectively captures semantic and syntactic relationships between words, where similar meanings or contextual uses result in similar vector positions. By reducing dimensionality while preserving these relationships, Word2Vec enhances the quality of textual analysis [22], [23]. In our study, we experimented with various vector dimensions and found that a dimension of 500 yielded the best results for detecting depression in the post-context.

2.4.3. Fine-tuned BERT

BERT is a natural language processing model that leverages a bidirectional transformer architecture to grasp the context of words within sentences. By analyzing text in both directions, BERT captures richer contextual information than traditional models. It is pre-trained on vast amounts of text and can be fine-tuned for specific tasks by adding tailored output layers. In our study, we fine-tuned BERT by incorporating additional output layers specifically designed for our post-classification task. This model was re-trained on our Reddit dataset to enhance its ability to capture the nuances of our corpus, ultimately optimizing its performance for depression detection [22], [24].

2.5. Machine learning algorithms

Machine learning algorithms are crucial for classifying data in natural language processing. They facilitate the identification and modeling of intricate relationships between input features and target categories using training data. Our algorithm selection is based on their strong reputation and widespread adoption within the scientific community for text classification tasks. These algorithms are known for their effectiveness across diverse contexts and their capacity to manage complex datasets, making them ideal for our study [25]–[27]. In our research, we employed the following techniques:

2.5.1. Decision tree

The decision tree (DT) algorithm is a classification model that splits data into subsets based on the values of specific features. Each internal node represents a decision based on a feature, while each leaf node corresponds to a target class. This algorithm is highly valued for its straightforward interpretability and capacity to capture complex relationships between features. It is clear decision-making process makes it particularly effective for classification tasks that require transparent and efficient results [25]–[27].

2.5.2. Logistic regression

Logistic regression (LR) is a classification algorithm that models the probability of an instance belonging to a particular class using a logistic function. This model is highly regarded for its simplicity, speed, and effectiveness in binary classification tasks, delivering results that are easy to interpret. Moreover, its ability to output probabilities alongside classifications allows for a more nuanced assessment of prediction uncertainties, enhancing decision-making in various applications [25]–[27].

2.5.3. Random forest

Random forest (RF) is a classification algorithm that enhances prediction accuracy and robustness by combining multiple decision trees. Each tree in the forest is constructed using a random subset of the data and features, and the final prediction is obtained by aggregating the predictions of all the trees. This ensemble approach mitigates the risk of overfitting and improves overall performance by leveraging the diversity among the individual trees, making it a powerful tool for complex classification tasks [25]–[27].

2.5.4. Gradient boosting

Gradient boosting (GB) is a classification algorithm that constructs a robust predictive model by sequentially combining multiple weak learners, typically decision trees. The algorithm works iteratively, where each new tree is trained to correct the errors made by the previous trees by focusing on the residuals. This process of iterative refinement reduces bias and enhances the overall accuracy of the model, making gradient boosting particularly effective in handling complex classification tasks with high precision [25]–[27].

2.5.5. Support vector machine

The support vector machine (SVM) is a powerful classification algorithm that aims to find the optimal hyperplane that separates different classes of data with the largest margin. It excels with complex, high-dimensional datasets and utilizes the kernel trick to handle non-linear problems by projecting data into a higher-dimensional space, thereby making them linearly separable. This ability to handle non-linearity effectively enhances its performance across diverse classification tasks [25]–[27].

2.6. Hyperparameter optimization using genetic algorithms

Hyperparameter optimization is essential for enhancing machine learning model performance by fine-tuning algorithm parameters to maximize accuracy and generalization [28]. In this study, we employed the HypONIC library, which implements a genetic algorithm (GA) inspired by natural selection to efficiently explore complex, high-dimensional search spaces. Unlike Grid Search, which exhaustively evaluates all possible parameter combinations and becomes computationally expensive in high-dimensional settings, or Bayesian optimization, which relies on probabilistic modeling and may struggle with highly non-linear search spaces, GA iteratively refines configurations through evolutionary operations such as selection, crossover, and mutation, allowing for adaptive exploration and the avoidance of local minima [29]. The GA was configured with a population size of 10 and was executed over 5 generations to balance computational efficiency and model optimization. A mutation rate of 0.1 was applied to introduce controlled variability, preventing premature convergence to suboptimal solutions, while a crossover rate of 0.9 facilitated the combination of high-performing configurations to enhance exploration. To ensure robust selection, a tournament selection strategy was employed, where the best-performing individuals were retained for the next generation, and elitism was enabled to preserve top solutions across iterations. By leveraging evolutionary principles, GA enables a more efficient and flexible hyperparameter search, reducing computational costs while improving model performance and stability [10], [30].

3. RESULTS AND DISCUSSION

We implemented our experimental framework on Google Colab's cloud platform, leveraging its 16 GB GPU acceleration to handle the computational workload. The technical stack utilized Python alongside essential machine learning libraries: Pandas for data manipulation, Scikit-learn for traditional algorithms, TensorFlow for neural network implementation, natural language toolkit (NLTK) for text preprocessing, Gensim for word embeddings, Transformers for BERT-based architectures, and HypONIC for hyperparameter optimization. While transformer models like BERT typically demand significant computational resources, our dataset of 7,731 posts allowed for efficient fine-tuning, completing training in practical timeframes without requiring specialized hardware. For comprehensive evaluation, we employed multiple performance metrics: accuracy, precision, recall, and F1-score [31], [32], which collectively assess different aspects of model effectiveness. The complete experimental results, presented in Tables 1-6, systematically compare the impact of various vectorization approaches (TF-IDF, Word2Vec, BERT) and

genetic algorithm optimization across our five machine learning architectures. This balanced approach between computational efficiency and methodological rigor demonstrates that our solution can be realistically deployed on standard cloud infrastructure, making it accessible for practical applications while maintaining robust performance standards.

Table 1. TF-IDF

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	90.69%	90.91%	90.31%	90.61%
Logistic regression (LR)	**94.96%**	98.06%	92.67%	**95.29%**
Random forest (RF)	85.46%	96.39%	73.30%	83.27%
Gradient booting (GB)	93.34%	**98.86%**	90.71%	94.61%
Support vector machine (SVM)	75.95%	68.22%	**96.07%**	79.78%

Table 2. Word2Vec

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	82.29%	81.45%	83.90%	82.66%
Logistic regression (LR)	86.30%	81.36%	93.72%	87.10%
Random forest (RF)	84.68%	84.26%	83.38%	83.82%
Gradient booting (GB)	**88.62%**	**89.57%**	89.92%	**89.75%**
Support vector machine (SVM)	87.59%	82.95%	**94.24%**	88.24%

Table 3. Fine-tuned BERT

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	86.88%	86.86%	86.52%	86.69%
Logistic regression (LR)	**95.41%**	96.39%	**94.24%**	**95.30%**
Random forest (RF)	88.69%	**96.67%**	79.84%	87.46%
Gradient booting (GB)	94.44%	97.08%	91.49%	94.20%
Support vector machine (SVM)	95.09%	95.87%	94.11%	94.98%

Table 4. TF-IDF + Genetic algorithm

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	92.31%	95.10%	89.01%	91.95%
Logistic regression (LR)	95.48%	95.25%	**94.50%**	**94.88%**
Random forest (RF)	**95.60%**	**98.15%**	76.31%	85.86%
Gradient booting (GB)	94.57%	96.20%	92.67%	94.40%
Support vector machine (SVM)	95.41%	94.85%	93.98%	94.41%

Table 5. Word2Vec + Genetic algorithm

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	82.61%	81.09%	83.64%	82.35%
Logistic regression (LR)	89.79%	85.56%	**95.42%**	90.22%
Random forest (RF)	86.17%	85.90%	86.13%	86.01%
Gradient booting (GB)	90.50%	90.54%	90.18%	90.36%
Support vector machine (SVM)	**90.89%**	**91.59%**	89.79%	**90.68%**

Table 6. Fine-tuned BERT + Genetic algorithm

	Accuracy	Precision	Recall	F1-score
Decision tree (DT)	87.27%	87.55%	86.52%	87.03%
Logistic regression (LR)	95.60%	96.52%	94.50%	95.50%
Random forest (RF)	89.46%	96.88%	81.28%	88.40%
Gradient booting (GB)	94.96%	**97.64%**	92.02%	94.74%
Support vector machine (SVM)	**95.99%**	96.68%	**95.16%**	**95.91%**

For the dataset vectorized using TF-IDF, as shown in Table 1, logistic regression delivered the best overall performance with an accuracy of 94.96% and an F1-score of 95.29%. Gradient boosting achieved the highest precision at 98.86%, while SVM excelled in recall with 96.07%. For the Word2Vec vectorization, detailed in Table 2, gradient boosting showed strong performance with an accuracy of 88.62%, precision of 89.57%, and an F1-score of 89.75%. SVM again demonstrated superior recall with a score of 94.24%. Lastly,

with the fine-tuned BERT model, as presented in Table 3, logistic regression led with an accuracy of 95.41%, recall of 94.24%, and an F1-score of 95.30%, while gradient boosting achieved the highest precision at 97.08%.

Following hyperparameter optimization with the genetic algorithm, significant performance improvements were observed for the TF-IDF model (Table 4), with a notable 19% enhancement for SVM (C: 6.42, gamma: 'scale', kernel: 'linear', degree: 4). Random forest (n_estimators: 40, criterion: 'gini', min_samples_split: 9, max_depth: 7) achieved the highest accuracy (95.60%) and precision (98.15%), while logistic regression excelled in recall (94.50%) and F1-score (94.88%). For the Word2Vec model (Table 5), optimization resulted in an average performance boost of 2%. SVM (C: 9.62, gamma: 'auto', kernel: 'linear', degree: 3) demonstrated the best accuracy (90.89%), precision (91.59%), and F1-score (90.68%), while logistic regression (penalty: 'l2', C: 870.15, max_iter: 180, solver: 'newton-cg') achieved the highest recall (95.42%). With the fine-tuned BERT model (Table 6), hyperparameter optimization led to an average improvement of 1%. SVM (C: 9.67, gamma: 'auto', kernel: 'linear', degree: 4) performed best in accuracy (95.99%), recall (95.16%), and F1-score (95.91%), while gradient boosting (loss: 'log_loss', criterion: 'friedman_mse', n_estimators: 70, max_depth: 7, min_samples_split: 4) achieved the highest precision (97.64%). To ensure the robustness of these results, we employed k-fold cross-validation, which mitigates performance variations by evaluating models across multiple training and testing splits. Given this rigorous validation approach, we did not perform additional statistical significance tests such as ANOVA or t-tests, which are more suited for comparing models trained on independent datasets rather than those assessed using cross-validation. Figure 2 provides a graphical overview of the performance of various models, offering a clear and intuitive visualization of the results.

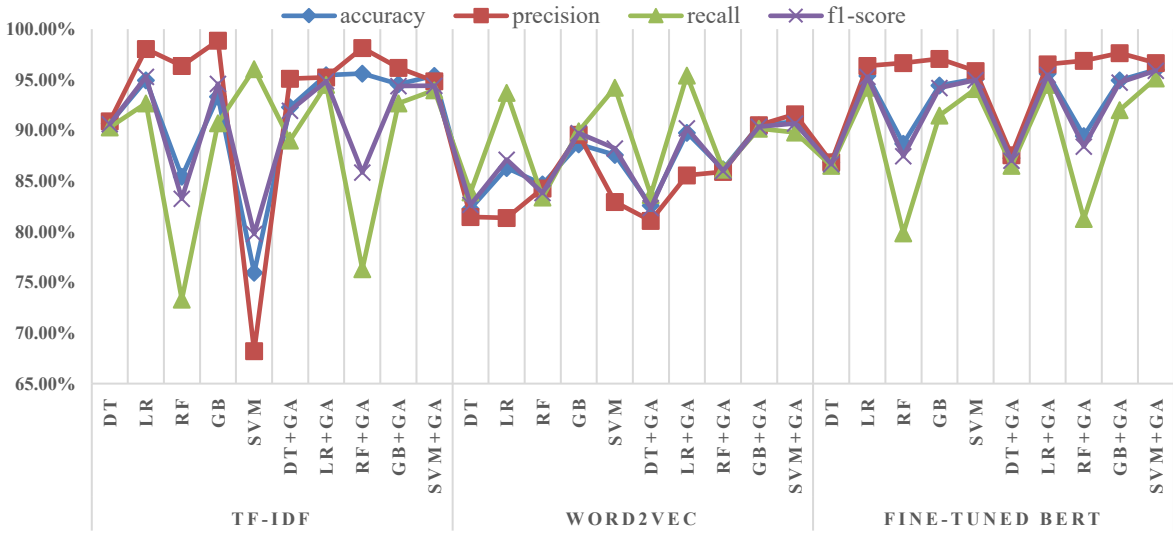

Figure 2. Metrics analysis for different algorithms

The results of our study indicate that the SVM model, when combined with a fine-tuned BERT and optimized using a genetic algorithm, achieved the highest overall performance. It attained an accuracy of 95.99%, a recall of 95.16%, and an F1-score of 95.91%. This exceptional performance of the SVM is attributed to its capacity to optimize the margin between classes through precise decision hyperplanes, thereby enhancing its accuracy in classifying positive examples. The fine-tuned BERT model was pivotal in achieving these results. Its adaptation to our dataset allowed it to effectively capture contextual nuances and complex relationships in Reddit comments, leading to significant performance improvements. Additionally, TF-IDF outperformed Word2Vec in our study. This is due to TF-IDF's effectiveness in highlighting term importance in short, specific texts like Reddit comments, whereas Word2Vec, which excels in capturing deep semantic relationships, was not as well-suited to our dataset. The lack of specific optimization for Word2Vec likely constrained its ability to capture the essential contextual nuances for accurate classification, emphasizing the need for vectorization methods to be tailored to the specific context of the data. Logistic regression also delivered impressive results, achieving an accuracy of 95.60% and an F1-score of 95.50%, highlighting its effectiveness in capturing linear relationships. Gradient boosting excelled in precision with a score of 97.64%, showcasing its strength in aggregating weak trees to correct classification errors. The

genetic algorithm played a crucial role in optimizing hyperparameters by efficiently exploring parameter spaces and significantly enhancing model performance. This method allowed us to fine-tune parameters effectively, achieving optimal configurations and the best possible results.

Comparing our results with those of previous studies, such as [12], [13], we find similarities in the application of machine learning techniques for depression detection. These studies also highlight the effectiveness of models like SVM in achieving robust results. For instance, [14] reported that SVMs achieved an accuracy of 77.12% and an F1-score of 77%. In another case, [15] observed that Random Forest performed well with an accuracy of 84.99% in analyzing social network posts, although their results were lower than those obtained in our study. In addition, [16] utilized a hybrid CNN and BiLSTM model, achieving an accuracy of 94.28% on depression-related tweets. [17] demonstrated that SVMs could reach 85% accuracy on tweets using various classifiers. Furthermore, [18] validated deep learning techniques like BERT, achieving an AUC of 93% for depression detection in medical records, although their model's performance varied depending on the context.

Our study sets itself apart by incorporating a genetic algorithm for hyperparameter optimization, a methodology seldom explored in previous research on depression detection. By fine-tuning model configurations, this approach led to 95.99% accuracy, 95.16% recall, and a 95.91% F1-score with the fine-tuned SVM and BERT model. Prior studies primarily relied on conventional tuning strategies such as grid search and random search, which lack the adaptive capabilities of genetic optimization [17], [19]. The integration of BERT fine-tuning with evolutionary optimization significantly enhances classification effectiveness, making it a promising avenue for real-world deployment. Beyond methodological advancements, our approach is highly adaptable for integration into mental health platforms and social media monitoring tools aimed at early depression detection. Potential implementations include API-based integration with health applications like Talkspace and Woebot, embedding within electronic health record systems, or automating preliminary screenings in clinical workflows. Future research could extend this work by incorporating multimodal analysis—such as speech sentiment detection and facial expression recognition—to improve diagnostic accuracy. Ethical considerations, including data privacy, informed consent, and algorithmic transparency, remain central to ensuring compliance with regulatory frameworks and minimizing bias in real-world applications. To systematically compare existing approaches with our methodology, Table 7 summarizes key studies in depression detection from online content, detailing their datasets, feature extraction techniques, and performance metrics. This synthesis highlights both the progression of technical approaches and the persistent gaps our study addresses, particularly through genetic algorithm optimization and BERT fine-tuning.

Table 7. Related work comparison

Study	Dataset source	Sample size	Feature extraction	Model(s) used	Best performance metrics	Key limitations
Arachchige *et al.* [12]	Reddit/Facebook/ Twitter forums	29 studies	Semantic patterns + behavioral markers	Various ML algorithms (SVM, LR, RF)	qualitative analysis	Cultural/language barriers, self-diagnosis errors, user anonymity
Glaz *et al.* [13]	Multi-platform forums, medical DBs, social media	58 studies	TF-IDF, LIWC, n-grams	Various ML algorithms (SVM, LR, RF)	review	Method heterogeneity, Cohort bias, Ethical limits
Jain *et al.* [14]	Reddit	60,000 records	Bag-of-Words, TF-IDF	SVM, NB, LR, RF	Acc: 77.12%, F1: 77%	Limited platform diversity
Saifullah *et al.* [15]	YouTube	4,862 comments	Count-Vectorization, TF-IDF	RF, LR, DT, SVM, KNN, NB	Acc: 84.99%	No deep learning, limited platform diversity
Kour *et al.* [16]	Twitter	2,558, 5304 records	CNN feature extraction + BiLSTM sequencing	CNN-BiLSTM	Acc: 94.28%	Seq-length issues, Platform bias, Missed comorbidities
Chereddy *et al.* [17]	Twitter	3,000 posts	TF-IDF	SVM, NB	Acc: 85%	Limited platform diversity
Chen *et al.* [18]	Medical records (Mandarin)	22,355 records	BERT embeddings	BERT-CNN	AUC: 93%	Military/civilian performance gaps
Subramanian *et al.* [19]	Multi-Social media platforms	-	FastText, n-grams	SVM, BERT, CNN	Acc: 80%	No severity grades, Label bias
Our Study	Reddit	7,731 posts	BERT, Word2Vec, TF-IDF	SVM (GA-optimized) + BERT	Acc: 95.99%	Platform bias, missing demographic analysis, sarcasm/ self-deprecation

4. LIMITATIONS

While our study advances depression detection in social media posts, several limitations must be acknowledged. The dataset, consisting solely of Reddit posts, may limit generalizability to other platforms, and the absence of demographic attributes like age and gender restricts bias analysis. Additionally, linguistic ambiguity remains a challenge, as sarcastic or self-deprecating statements may be misclassified, leading to false positives, while false negatives could overlook individuals in need. To address these risks, refining classification thresholds, incorporating human oversight, and positioning these models as supportive tools rather than standalone diagnostic solutions are crucial steps. From a computational perspective, transformer-based models typically require significant resources. However, our experiments on Google Colab with a 16 GB GPU demonstrated that fine-tuning BERT on this dataset was feasible within a reasonable time frame, suggesting that cloud-based deployment is a viable option. Future studies should further explore efficiency trade-offs, particularly when scaling to larger datasets or real-time applications. Enhancing model interpretability is another important direction. Techniques such as SHAP and LIME could provide transparent explanations for predictions, which is especially relevant as future research integrates additional user attributes to improve classification accuracy. Moreover, extending this framework beyond depression detection to other mental health conditions, such as anxiety or bipolar disorder, could broaden its applicability. Finally, while we optimized BERT using a genetic algorithm, future research could investigate alternative transformer architectures such as RoBERTa or GPT, as well as compare GA with Bayesian optimization or grid search for hyperparameter tuning. Additionally, validating the model on clinically annotated datasets and incorporating multimodal data—such as speech or behavioral cues—could further strengthen its effectiveness in real-world mental health assessment.

5. CONCLUSION

This study demonstrates the effectiveness of machine learning techniques, particularly SVM with fine-tuned BERT, in detecting depressive content in online comments. By incorporating hyperparameter optimization through genetic algorithms, we achieved impressive results: 95.99% accuracy, 95.16% recall, and 95.91% F1-score. This advanced approach significantly enhances predictive performance compared to previous research, highlighting how sophisticated model tuning and BERT's contextual understanding can advance tools for detecting depressive disorders. The implications of our research extend beyond the immediate findings. Integrating BERT fine-tuning with genetic algorithm optimization establishes a robust framework for developing more accurate and reliable predictive tools. This approach has the potential to revolutionize the diagnosis and management of depressive disorders. Future research should validate these results across diverse datasets and clinical settings to ensure broader applicability and robustness. Additionally, integrating these methods with electronic health records and real-time patient monitoring systems could further enhance their utility. Ultimately, our study advances predictive medicine by providing a practical approach to detecting depressive content, paving the way for personalized treatment strategies and improved patient management, with significant implications for both research and clinical practice.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abd Allah Aouragh	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Mohamed Bahaj	✓	✓			✓		✓			✓		✓	✓	
Fouad Toufik	✓	✓			✓		✓			✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, AA, upon reasonable request.

REFERENCES

[1] M. Maj *et al.*, "The clinical characterization of the adult patient with depression aimed at personalization of management," *World Psychiatry*, vol. 19, no. 3, pp. 269–293, 2020, doi: 10.1002/wps.20771.

[2] WHO, "Depressive disorder (depression)," *The ECPH Encyclopedia of Psychology*. 2025, Accessed: Aug. 26, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] L. Cui *et al.*, "Major depressive disorder: hypothesis, mechanism, prevention and treatment," *Signal Transduction and Targeted Therapy*, vol. 9, no. 1, pp. 1–32, 2024, doi: 10.1038/s41392-024-01738-y.

[4] P. Cuijpers, A. Stringaris, and M. Wolpert, "Treatment outcomes for depression: challenges and opportunities," *The Lancet Psychiatry*, vol. 7, no. 11, pp. 925–927, 2020, doi: 10.1016/S2215-0366(20)30036-5.

[5] K. A. McLaughlin, "The public health impact of major depression: a call for interdisciplinary prevention efforts," *Prevention Science*, vol. 12, no. 4, pp. 361–371, 2011, doi: 10.1007/s11121-011-0231-8.

[6] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–13, 2022, doi: 10.1038/s41746-022-00589-7.

[7] B. G. Teferra *et al.*, "Screening for depression using natural language processing: literature review," *Interactive Journal of Medical Research*, vol. 13, p. e55067, 2024, doi: 10.2196/55067.

[8] D. D. DeSouza, J. Robin, M. Gumus, and A. Yeung, "Natural language processing as an emerging tool to detect late-life depression," *Frontiers in Psychiatry*, vol. 12, p. 719125, 2021, doi: 10.3389/fpsyt.2021.719125.

[9] P. Shweta and K. Adhiya, "Comparative study of feature engineering for automated short answer grading," in *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022*, 2022, pp. 594–597, doi: 10.1109/AIC55036.2022.9848851.

[10] D. L. Shanthi and N. Chethan, "Genetic algorithm based hyper-parameter tuning to improve the performance of machine learning models," *SN Computer Science*, vol. 4, no. 2, p. 119, 2023, doi: 10.1007/s42979-022-01537-8.

[11] R. Jain, R. S. Rai, S. Jain, R. Ahluwalia, and J. Gupta, "Real time sentiment analysis of natural language using multimedia input," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 41021–41036, 2023, doi: 10.1007/s11042-023-15213-3.

[12] I. A. N. Arachchige, P. Sandanapitchai, and R. Weerasinghe, "Investigating machine learning & natural language processing techniques applied for predicting depression disorder from online support forums: A systematic literature review," *Information (Switzerland)*, vol. 12, no. 11, 2021, doi: 10.3390/info12110444.

[13] A. Le Glaz *et al.*, "Machine learning and natural language processing in mental health: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, 2021, doi: 10.2196/15708.

[14] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and suicide analysis using machine learning and NLP," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 12034, 2022, doi: 10.1088/1742-6596/2161/1/012034.

[15] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *Jurnal Informatika*, vol. 15, no. 1, p. 45, 2021, doi: 10.26555/jifo.v15i1.a20111.

[16] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 23649–23685, 2022, doi: 10.1007/s11042-022-12648-y.

[17] S. Chereddy, K. Geetha, and A. G. Sreedevi, "Tweeting the blues: leveraging NLP and classification models for depression detection," in *Proceedings - 2nd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2024*, 2024, pp. 875–880, doi: 10.1109/InCACCT61598.2024.10550961.

[18] T.- Chen, H.- Chu, Y.- Tai, and S.- Yang, "Performances of depression detection through deep learning-based natural language processing to mandarin Chinese medical records: Comparison between civilian and military populations.," *Taiwanese Journal of Psychiatry*, vol. 36, no. 1, pp. 32–38, 2022.

[19] M. Subramanian, G. Raju, A. Sureshkumar, C. Anbarasu, K. S. Vadivel, and P. S. Nandhini, "From words to emotions: identifying depression through social media insights," *Communications in Computer and Information Science*, vol. 2046 CCIS, pp. 268–282, 2024, doi: 10.1007/978-3-031-58495-4_20.

[20] InFamousCoder, "Depression: Reddit dataset (cleaned)" *Kaggle*, 2024, Accessed: Aug. 26, 2024. [Online]. Available: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned.

[21] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Information Systems*, vol. 121, 2024, doi: 10.1016/j.is.2023.102342.

[22] J. Ravi and S. Kulkarni, "Text embedding techniques for efficient clustering of twitter data," *Evolutionary Intelligence*, vol. 16, no. 5, pp. 1667–1677, 2023, doi: 10.1007/s12065-023-00825-3.

[23] L. Jiaxin, "Text vectorization in sentiment analysis: A comparative study of TF-IDF and Word2Vec from Amazon fine food reviews," *ITM Web of Conferences*, vol. 70, p. 3001, 2025.

[24] P. Asmi and M. S. Sanaj, "Toxic speech classification via deep learning using combined features from BERT & FastText embedding," *International Journal of Engineering Research & Technology (IJERT), ICCIDT - 2021 Conference Proceedings*, vol. 9, no. 7, pp. 68–71, 2021, doi: 10.17577/IJERTCONV9IS07016.

[25] A. Ali Raza *et al.*, "Review to unfold the role of machine learning algorithms in natural language processing," *Journal of Policy Research*, vol. 9, no. 4, pp. 152–162, Dec. 2023, doi: 10.61506/02.00136.

[26] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-1004-8.

[27] T. Sood, P. Garg, H. Pannu, and S. Dullat, "Machine learning algorithms and NLP for anxiety and depression detection using Twitter data," in *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and*

Biomedical Health Informatics, IC3ECSBHI 2025, 2025, pp. 450–455, doi: 10.1109/IC3ECSBHI63591.2025.10990488.
[28] Monica and P. Agrawal, "A survey on hyperparameter optimization of machine learning models," in *2024 2nd International Conference on Disruptive Technologies, ICDT 2024*, 2024, pp. 11–15, doi: 10.1109/ICDT61202.2024.10489732.
[29] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, 2021, doi: 10.3390/informatics8040079.
[30] H. Alibrahim and S. A. Ludwig, "Hyperparameter optimization: comparing genetic algorithm against grid search and bayesian optimization," in *2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings*, 2021, pp. 1551–1559, doi: 10.1109/CEC45853.2021.9504761.
[31] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLoS ONE*, vol. 18, no. 10 October, 2023, doi: 10.1371/journal.pone.0291908.
[32] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-56706-x.

BIOGRAPHIES OF AUTHORS

Abd Allah Aouragh is currently pursuing his Ph.D. at the MIET Laboratory, Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco. His research is focused on advancing medical diagnosis support systems, with a particular emphasis on leveraging machine learning, deep learning, and computer vision techniques to develop innovative solutions. For inquiries or collaboration opportunities, Abd Allah can be reached via email at abdallahaouragh@gmail.com.

Mohamed Bahaj received his Ph.D. in mathematics and computer science from the University Hassan 1st, Morocco, and currently serves as a Full Professor in the Department of Mathematics and Computer Sciences at the University Hassan 1st, Faculty of Sciences and Technology, Settat, Morocco. With a robust academic background, he has contributed over 60 peer-reviewed papers, spanning areas such as intelligent systems, ontologies engineering, partial and differential equations, numerical analysis, and scientific computing. He has provided valuable peer review services for various journals and mentored several Ph.D. students in computer sciences and mathematics. He actively engages in workshops, seminars, and academic forums to enhance teaching methodologies and research practices. For inquiries, he can be contacted via email at mohamedbahaj@gmail.com.

Fouad Toufik received his Ph.D. in computer science from the University Hassan 1st, Settat, Morocco. He currently holds the position of Professor of Computer Sciences at the Higher School of Technology SALE, Mohammed V University, Morocco. With expertise in artificial intelligence, big data, and database architectures, his research interests lie at the intersection of these fields. His academic pursuits aim to advance knowledge and innovation in these areas, contributing to the development of cutting-edge technologies. For inquiries or collaboration opportunities, he can be contacted via email at toufik.fouad@gmail.com.