DETECTION OF DEPRESSION STATUS FROM TEXT USING NLP APPROACH: COMPARATIVE PERFORMANCE OF DIFFERENT ENSEMBLE ALGORITHMS

Doç. Dr. Zülfikar ASLAN

Gaziantep University, Gaziantep, Turkey, zulfikaraslan27@gmail.com

ABSTRACT

This study aims to compare the performance of different ensemble learning algorithms using Natural Language Processing (NLP) approach for the detection of depression status from text. Depression is a common mental health problem worldwide and early diagnosis and intervention are of critical importance. A dataset consisting of 124,017 tweets collected between 2019-2020 was used in the study. These tweets were classified according to their depressive and non-depressive content. Five different ensemble learning algorithms, namely Random Forest, AdaBoost, Gradient Boosting, XGBoost and Voting Classifier, were applied in the study. The performance of the models was evaluated using various metrics such as accuracy, precision, recall and F1-score. In addition, ROC curves and learning curves were analyzed. The results revealed that all ensemble learning algorithms showed high performance, but Random Forest gave the best results in all metrics. Voting Classifier performed the second best, while Gradient Boosting performed relatively poorly. This study shows that ensemble learning algorithms are effective in detecting depression from text, and Random Forest in particular can be used as a potential screening tool in this area. The findings may contribute to the development of technology-supported approaches for early diagnosis and intervention in the field of mental health.

Keywords: Depression Detection, Natural Language Processing (NLP), Ensemble Learning Algorithms, Social Media Analysis

1. INTRODUCTION

Depression is a serious mental health problem that affects millions of people worldwide. According to the World Health Organization, more than 300 million people suffer from depression globally [1]. This common and serious condition requires early diagnosis and intervention. In recent years, with the development of technology, new methods for the detection of depression have begun to be investigated. Among these methods, the use of Natural Language Processing (NLP) techniques stands out.

NLP is a branch of artificial intelligence that deals with the development of computer systems capable of understanding, processing and producing the language that people use in daily life. NLP techniques can be used to analyze the emotional states of individuals from their written or spoken expressions. In this context, the detection of depression symptoms from textual data offers a promising approach for early diagnosis and intervention [2].

The aim of this study is to compare the performance of different ensemble algorithms for the detection of depression from text using the NLP approach. Ensemble learning algorithms have the potential to make more powerful and reliable predictions by combining multiple machine learning models. There are several important advantages of using these algorithms:

1. Performance Improvement: Ensemble methods generally outperform a single model. By combining the strengths of different models, more accurate and reliable results can be obtained [3].

2. Generalization Ability: Ensemble models can generalize better across different datasets and scenarios. This is especially important for complex and multidimensional problems such as depression detection [4].

3. Resistance to Overfitting: Ensemble methods reduce the risk of overfitting compared to a single model. This prevents the model from overfitting the training data, allowing it to perform better on new and unseen data [5].

4. Management of Uncertainty: Ensemble models can better manage uncertainty in predictions. This is especially critical in the healthcare field, where false positive or false negative results can have serious consequences [6].

This research aims to determine which ensemble algorithms are most effective in detecting depression and to contribute to the current knowledge in this field. The above-mentioned advantages of ensemble learning algorithms explain why these methods are preferred in depression detection.

Related Literature Studies

Studies on depression detection using NLP techniques have increased in recent years. In this section, we will review important and highly cited studies on the subject, especially focusing on ensemble learning algorithms.

NLP and Depression Detection

De Choudhury et al. (2013) conducted one of the pioneering studies on detecting depression symptoms using social media data [7]. The researchers analyzed the posts of Twitter users and examined the relationship between factors such as language use, mood, and social interaction with depression. The study showed that social media data can be used to detect depression.

Eichstaedt et al. (2018) tried to predict depression diagnosis using Facebook posts [8]. The researchers studied a large sample of 683 patients diagnosed with depression and

28,749 controls. The results showed that language use in Facebook posts was effective in predicting future depression diagnoses.

Ensemble Learning Algorithms and Depression Detection

Sagi and Rokach (2018) conducted a comprehensive review on the use of ensemble learning methods in the healthcare domain [9]. This study discussed how different ensemble algorithms can be applied to healthcare data and their potential benefits. The researchers highlighted that ensemble methods outperform a single model, especially in complex healthcare problems. Trotzek et al. (2020) compared deep learning and ensemble methods for depression detection [10]. The researchers tried to detect depression symptoms from social media data using different NLP techniques and model architectures. The results revealed that ensemble methods outperform a single model. Sau and Bhakta (2017) investigated the performance of ensemble learning algorithms in text classification problems [11]. The study compared popular ensemble algorithms such as Random Forest, Gradient Boosting, and XGBoost and evaluated how these methods perform in sentiment analysis and text classification tasks. Orabi et al. (2018) conducted a comprehensive study combining deep learning and ensemble methods for depression detection [12]. The researchers developed an ensemble approach based on deep learning models such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). In the study, experiments on social media data showed that the proposed ensemble model achieved higher accuracy and F1 score than a single deep learning model. This study revealed that combining deep learning and ensemble methods can be a powerful approach for depression detection. Burdisso et al. (2019) presented a new ensemble learning approach for depression detection in time series data [13]. The proposed method aimed to better capture the temporal changes of depression symptoms by combining the results of multiple models operating in different time windows.

This literature review provides a summary of existing studies in the field of depression detection using NLP techniques and ensemble learning algorithms. It is seen that ensemble learning algorithms show promising results in depression detection. However, there are still gaps to be investigated in the comparative performance of different ensemble algorithms. This study aims to fill this gap and determine which ensemble algorithms are most effective in detecting depression.

2. MATERIAL and METHODS

Proposed Method

In this study, a machine learning approach is proposed for depression detection from text data. The proposed method includes data preparation, model training, and a comprehensive evaluation process. The main steps of the method are as follows:

1. Data Preparation and Preprocessing

The first step is to load the dataset and preprocess it. In this stage, the text data is cleaned, normalized, and made ready for analysis. Preprocessing steps may include processes such as removing unnecessary characters, converting to lowercase, and cleaning stop words.

2. Vectorization of Text Data

The preprocessed text data is converted into a numerical format that machine learning algorithms can process. Modern NLP techniques such as TF-IDF, Word2Vec, or BERT can be used for this process.

3. Splitting the Dataset

The vectorized dataset is split into training and test sets. This split provides an unbiased evaluation of the model's performance.

4. Definition of Machine Learning Algorithms

The machine learning algorithms to be used in the study are determined and configured. At this stage, combinations of deep learning models with classical algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forest can be considered.

5. Model Evaluation

The trained models are evaluated comprehensively using various metrics. This evaluation process includes the following steps:

- Creation of ROC (Receiver Operating Characteristic) curves
- Calculation of confusion matrices
- Drawing of learning curves
- Presentation of performance results in a table

6. Visualization of Results

The results obtained from the model evaluation are presented using various visualization techniques:

- Creation of feature discretization graph
- Drawing histograms and boxplots
- Displaying performance metrics as a heatmap

This comprehensive evaluation and visualization process allows comparing the performance of different models and determining the most effective approach.

The proposed method aims to provide a reliable and interpretable framework for text-based depression detection. The results of the study will reveal which algorithms and features are most effective in depression detection and will guide future research. Figure 1 shows the flow chart of the proposed study.

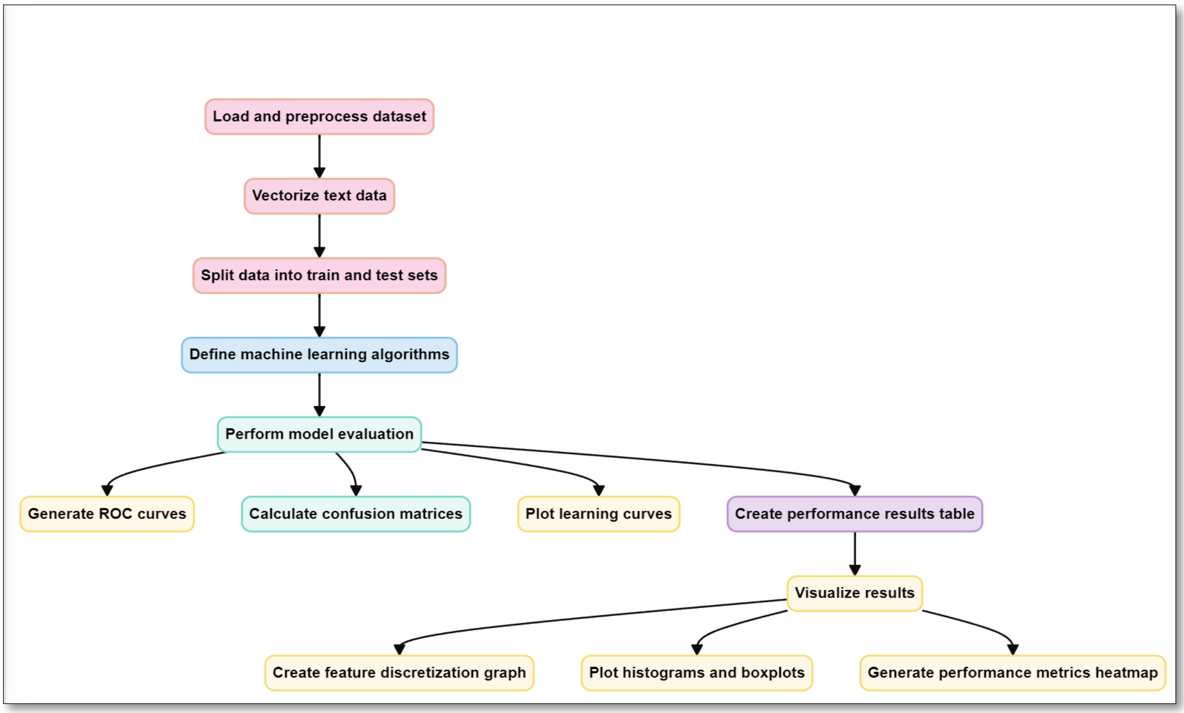

Load and preprocess dataset
Vectorize text data
Split data into train and test sets
Define machine learning algorithms
Perform model evaluation
Generate ROC curves
Calculate confusion matrices
Plot learning curves
Create performance results table
Visualize results
Create feature discretization graph
Plot histograms and boxplots
Generate performance metrics heatmap

Figure 1. Flow chart of the proposed method

Dataset

Tweets obtained from December 2019 to December 2020, mainly from India and the South Asian subcontinent, were classified according to their depressive and non-depressive content. Each tweet was assigned an emotional value using the Sentiment Scoring method, one of the text mining techniques. The tweets in question were examined with a specially designed natural language processing algorithm based on 250 negative and positive word lists obtained from the SentiWord database and various academic studies. The entire dataset contains 124017 records [14].

Ensemble Classifiers

Ensemble learning is a method of making more powerful predictions by combining multiple machine learning models. This approach aims to overcome the limitations of a single model and achieve higher generalization ability [15]. Ensemble methods have the potential to reduce overfitting and increase prediction accuracy, especially in complex and noisy data sets [16]. In this study, five different ensemble learning algorithms were used for depression detection: Random Forest, AdaBoost, Gradient Boosting, XGBoost and Voting Classifier. The reason for choosing each algorithm and its features are explained below:

1. Random Forest: Random Forest, developed by Breiman, is an ensemble method that combines many decision trees [17]. It was preferred because it performs effectively in high-dimensional data and can determine feature importance levels.

2. AdaBoost (Adaptive Boosting): AdaBoost, proposed by Freund and Schapire, transforms weak learners into a strong classifier [18]. It was chosen because it performs well in noisy data and is resistant to overfitting.

3. Gradient Boosting: Gradient Boosting, developed by Friedman, increases model performance by sequentially training weak learners [19]. It was preferred due to its high prediction accuracy and ability to determine variable importance levels.

4. XGBoost (Extreme Gradient Boosting): XGBoost, developed by Chen and Guestrin, is a faster and more effective version of Gradient Boosting [20]. It was selected due to its high performance on large data sets and regularization features.

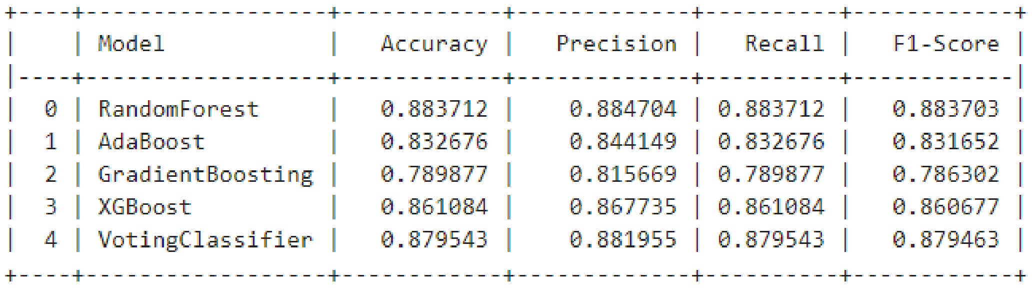
5. Voting Classifier: It is a meta-classifier that makes decisions by combining the predictions of different classifiers [4]. It was preferred because it combines the strengths of various models and makes more balanced and reliable predictions.

The use of these ensemble algorithms aims to achieve higher accuracy and reliability in depression detection. Considering the complex structure of text-based data and the diversity of depression symptoms, ensemble methods are expected to be effective in this challenging task [8].

In our study, the performances of these algorithms will be compared and which ensemble method is most effective in depression detection will be determined. This comparison will guide future studies and reveal the effectiveness of machine learning approaches in detecting depression.

3. EXPERIMENTAL RESULTS

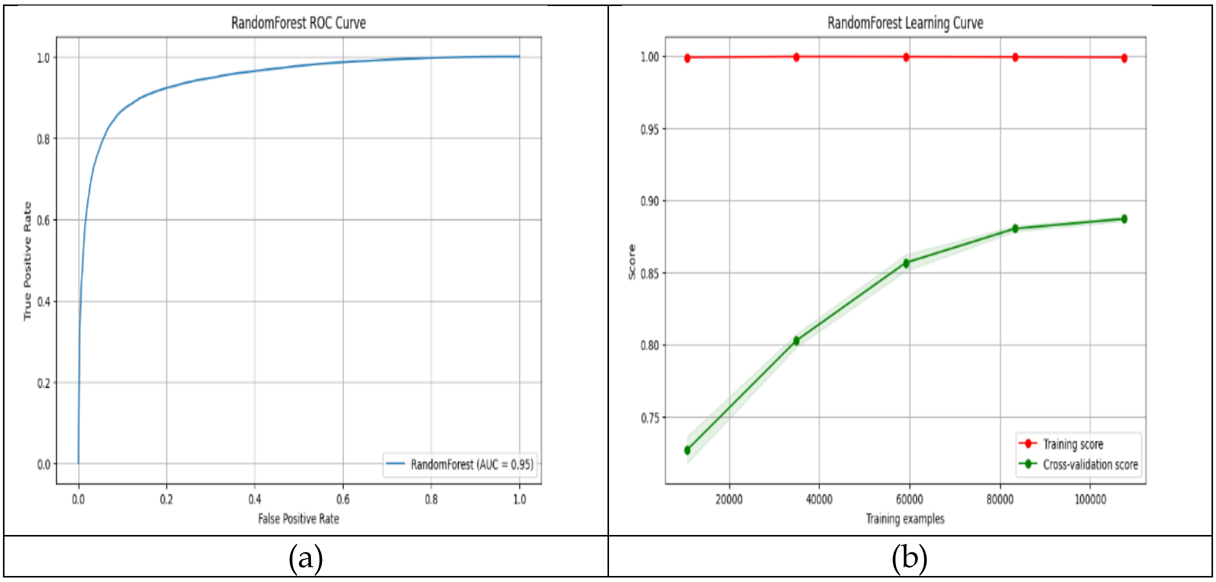
All experimental processes in the reviewed study were performed on the Colab platform using the Python programming language. To solve the problem of over-learning and under-learning, all classifiers were trained using 5-fold cross-validation. Figure 2 shows the evaluation metrics obtained as a result of the classification.



```
+----+------------------+------------+-------------+----------+------------+
|    | Model            |   Accuracy |   Precision |   Recall |   F1-Score |
|----+------------------+------------+-------------+----------+------------|
|  0 | RandomForest     |   0.883712 |    0.884704 | 0.883712 |   0.883703 |
|  1 | AdaBoost         |   0.832676 |    0.844149 | 0.832676 |   0.831652 |
|  2 | GradientBoosting |   0.789877 |    0.815669 | 0.789877 |   0.786302 |
|  3 | XGBoost          |   0.861084 |    0.867735 | 0.861084 |   0.860677 |
|  4 | VotingClassifier |   0.879543 |    0.881955 | 0.879543 |   0.879463 |
+----+------------------+------------+-------------+----------+------------+
```

Figure 2. Evaluation metrics calculated during the classification process.

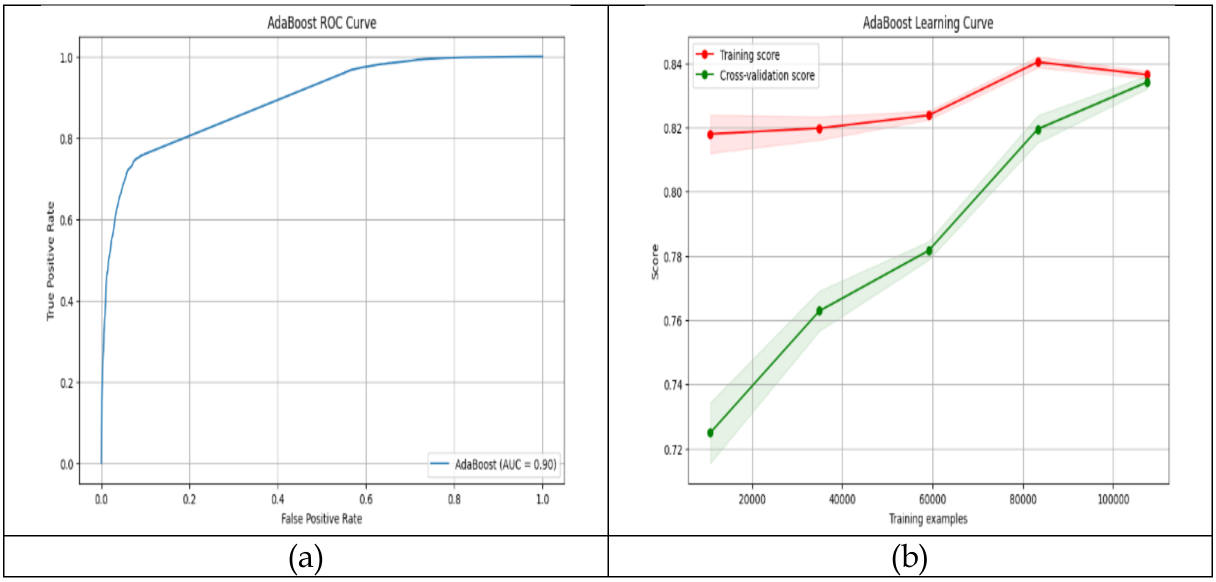
When Figure 2 is examined, the Random Forest model stands out as the most effective ensemble learning algorithm for this dataset and problem. Although other models also exhibit acceptable performance, additional improvements may be required, especially to improve the performance of Gradient Boosting.



(a) (b)

Figure 3. (a) ROC Curve and (b) Training Curve of Random Forest classifier.

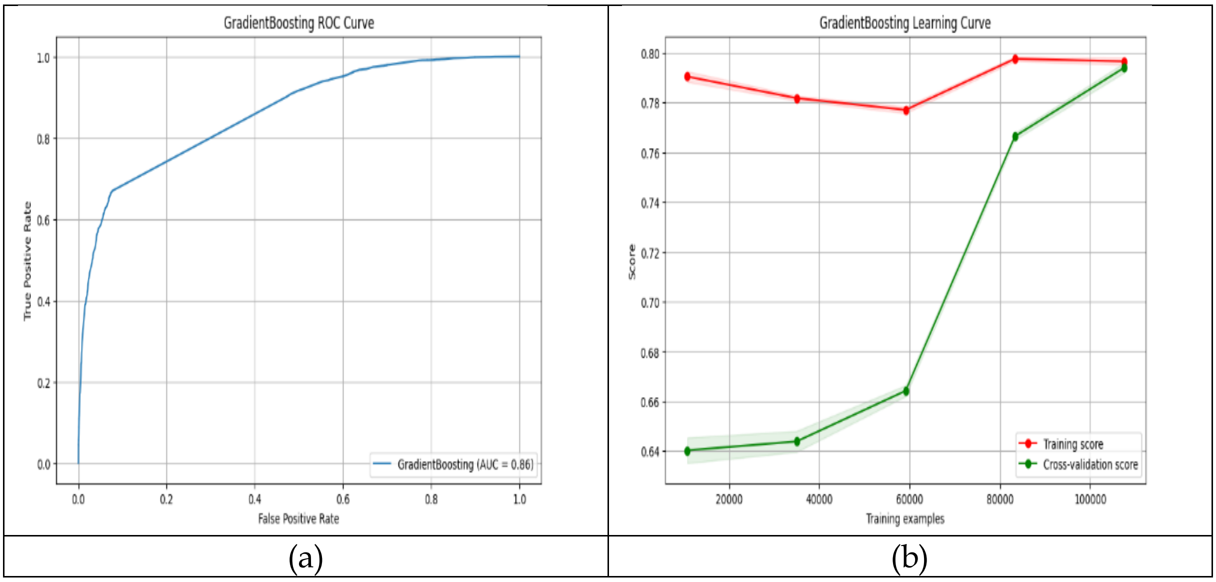
In Figure 3, the area under the ROC curve (AUC) value of the Random Forest model is 0.95. This value shows that the model performs very well and can distinguish positive and negative classes with high accuracy. In general, the Random Forest model performs excellently on the training data, while the cross-validation scores are also quite high. As the number of training examples increases, the cross-validation score also increases, indicating that the generalization ability of the model is strengthened. These findings obtained during the training of the model show that the model can classify with high accuracy and that its generalization ability is good when trained with sufficient data to avoid overfitting.



(a) (b)

Figure 4. (a) ROC curve and (b) Training curve of AdaBoost classifier.

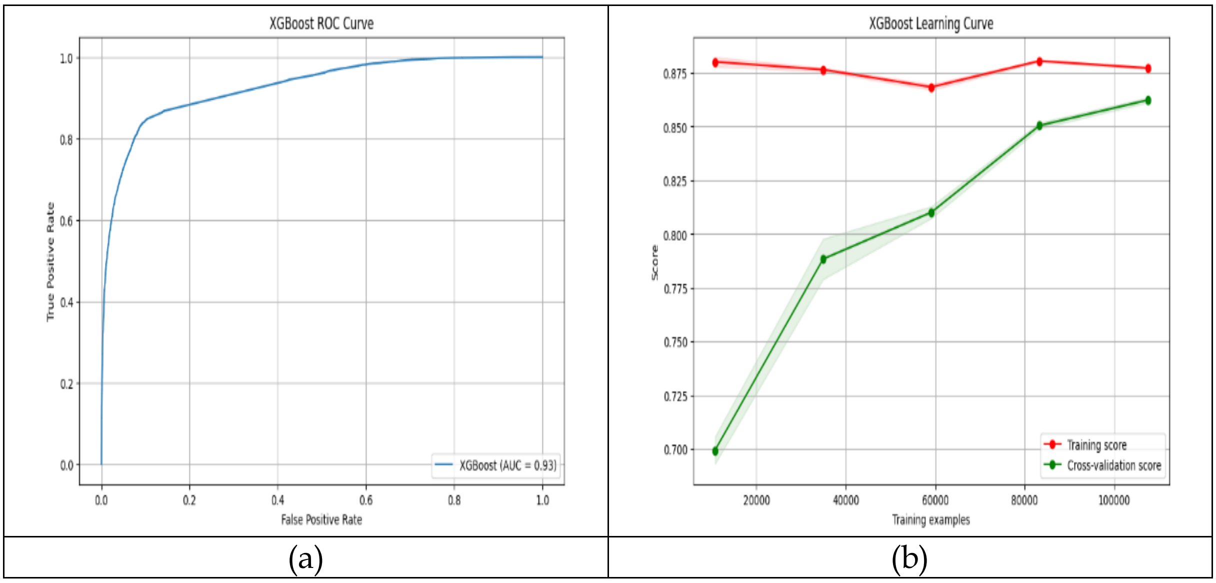
In Figure 4, the area under the ROC curve (AUC) of the AdaBoost model is 0.90. This value shows that the model performs quite well and can distinguish between positive and negative classes. The curve being close to the upper left corner indicates that the model has high sensitivity and low false positive rate. In general, the AdaBoost model performs well on the training data, while the cross-validation scores are also high and improve as the number of training examples increases. However, there is a difference between the training score and the cross-validation score; this difference may indicate that the model is slightly overfitting the training data. The increasing trend in the cross-validation scores indicates that the generalization ability of the model improves as the amount of data increases. As a result, it is seen that the AdaBoost model performs well in general, but its generalization ability can increase even more when trained with more data. Both the training and cross-validation performances of the model show that it has sufficient success in the classification task.



(a) (b)

Figure 5. (a) ROC curve and (b) Training curve of GradientBoosting classifier.

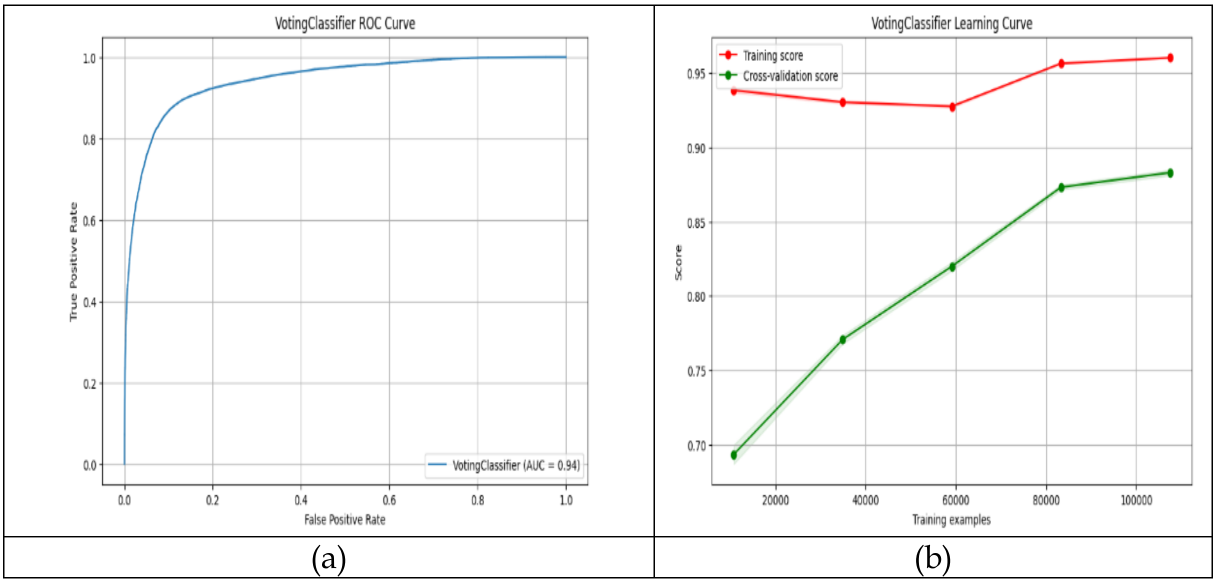
In general, when the ROC curve and learning curve in Figure 5 are examined, it can be said that the GradientBoosting classifier exhibits good performance and the generalization ability of the model increases with the increase in training data. This shows that the model provides satisfactory performance both on training data and on new, previously unseen data.



(a) (b)

Figure 6. (a) ROC curve and (b) Training curve of XGBoost classifier.

When the ROC curve and learning curve in Figure 6 are examined, it can be said that the XGBoost classifier exhibits very good performance and the generalization ability of the model increases with the increase in training data. This shows that the model provides satisfactory performance both on training data and on new, previously unseen data. The AUC value of XGBoost is 0.91, which shows that this model has a very strong discriminatory ability.



(a) (b)

Figure 7. (a) ROC curve and (b) Training curve of VotingClassifier classifier.

When the ROC curve and learning curve in Figure 7 are examined, it can be said that the VotingClassifier classifier exhibits very good performance and the generalization ability of the model increases with the increase in training data. This shows that the model provides satisfactory performance both on training data and on new, previously unseen data. The AUC value of VotingClassifier is 0.94, which shows that this model has a very strong discriminatory ability.

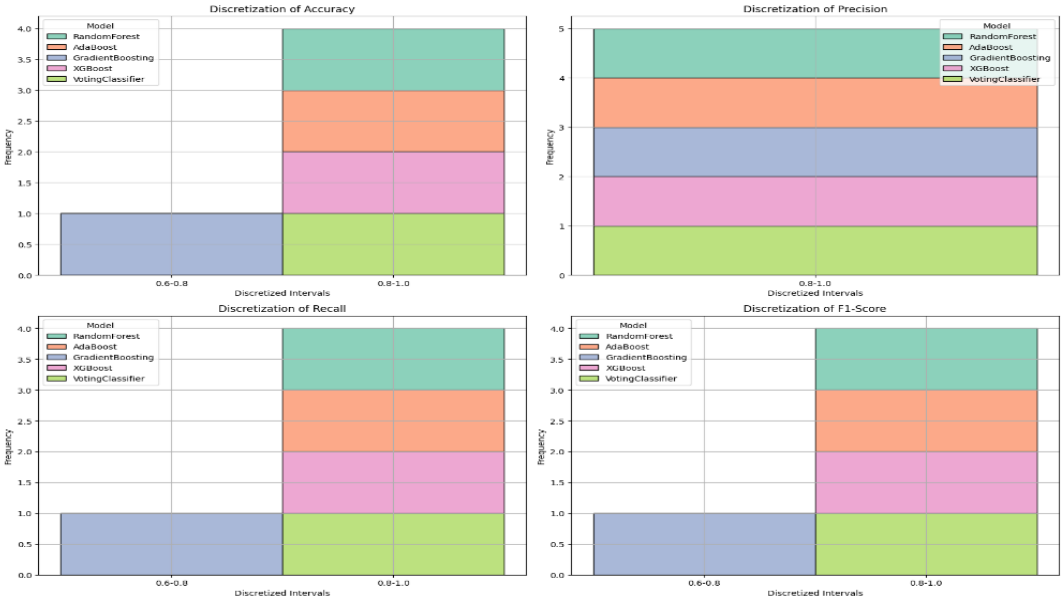


Figure 8. Various ensemble learning models for different evaluation metrics.

Figure 8 presents a comparative analysis of various machine learning models for four different performance metrics. These metrics appear as accuracy, precision, recall, and F1 score, respectively. Each graph shows two discrimination ranges (0.0-0.5 and 0.5-1.0) on the x-axis and frequency on the y-axis. Five different models are compared: RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier, XGBClassifier, and VotingClassifier.

General observations:

1. For all metrics, the performance of the models is generally concentrated in the range of 0.5-1.0, indicating that the models perform well overall.

2. RandomForestClassifier appears to have the highest frequency consistently across all metrics, especially in the range of 0.5-1.0.

3. VotingClassifier generally stands out as the second best performing model.

4. AdaBoostClassifier and GradientBoostingClassifier show moderate performance. 5. XGBClassifier appears to have the lowest frequency in most metrics, but still falls within the 0.5-1.0 range.

Metric-based evaluation:

- Accuracy: All models are concentrated in the range of 0.5-1.0, indicating good overall classification performance.

- Precision: More variation is observed among models, but still mostly in the range of 0.5-1.0.

- Recall: It shows a similar distribution to the accuracy metric, with high performance observed for all models.

- F1 Score: F1 score, which is the harmonic mean of precision and sensitivity, reflects the overall performance of the models and is again concentrated in the high range.

This analysis reveals that ensemble methods (especially RandomForest and VotingClassifier) show consistent and high performance on this particular problem. However, the relatively lower performance of XGBClassifier can potentially be improved by hyperparameter optimization or improvement of data preprocessing techniques.

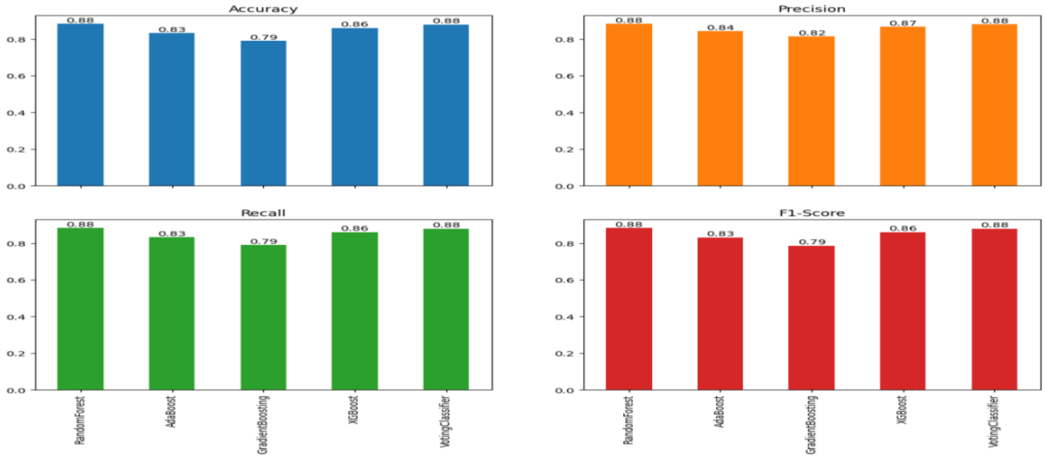


Figure 9. Performance of ensemble classifiers for each evaluation metric.

When Figure 9 is analyzed, all classifiers show very high performance. For most metrics, values above 0.80 are obtained, indicating that the models generally perform well. These results show that ensemble learning methods are quite effective for this classification problem. The superior performance of RandomForest suggests that this technique is particularly suitable for the dataset and problem structure. In future studies, it can be considered that the performance can be further improved by optimizing the model hyperparameters or trying different ensemble strategies.

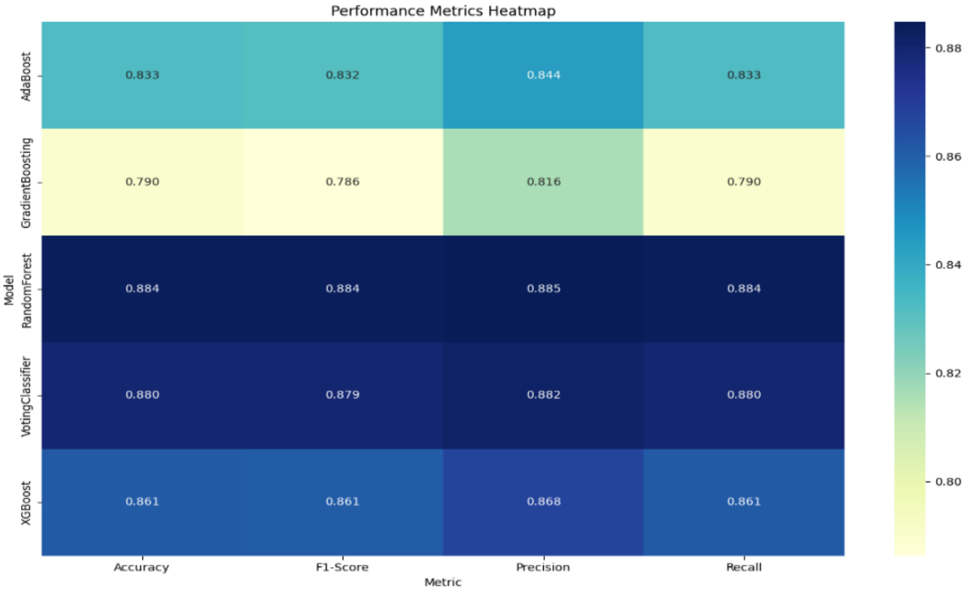


Figure 10. Evaluation metrics heatmap.

The heatmap presented in Figure 10 and the previous graph show that ensemble learning methods are generally effective for this classification problem, but RandomForest and VotingClassifier stand out in particular. This comparative analysis provides valuable information when performing model selection and fine-tuning.

CONCLUSIONS

In this study, the performance of different ensemble learning algorithms for the detection of depression from text was compared. As a result of the analyses performed using Random Forest, AdaBoost, Gradient Boosting, XGBoost and Voting Classifier algorithms, the following main findings were reached:

1. All ensemble learning algorithms showed high performance in the detection of depression from text. This shows that ensemble learning methods are suitable for such complex classification tasks.

2. Random Forest algorithm showed the highest performance in all evaluation metrics (accuracy, precision, sensitivity and F1-score). This shows that Random Forest is particularly effective for text-based depression detection.

3. Voting Classifier showed the second best performance overall. This shows that combining predictions from different models can be advantageous.

4. Gradient Boosting algorithm showed relatively lower performance compared to other methods. However, it is thought that this performance can be improved with hyperparameter optimization.

5. When the ROC curves and learning curves for all models were examined, it was observed that the generalization abilities of the models were good and the overfitting problem was minimal.

These results show that ensemble learning algorithms are an effective approach for the detection of depression from text. In particular, the superior performance of the Random Forest algorithm suggests that this method can be used as a potential screening tool in clinical applications. The following suggestions can be made for future studies:

1. Testing the models on larger and more diverse data sets.

2. Further improving the model performances with hyperparameter optimization and different feature engineering techniques.

3. Combining deep learning methods with ensemble learning algorithms.

4. Conducting studies to increase the explainability of the models.

5. Evaluating the practical applicability of these models in clinical settings.

In conclusion, this study has demonstrated the potential of ensemble learning algorithms for the detection of depression from text. These findings may contribute to the development of technology-supported approaches for early diagnosis and intervention in the field of mental health.

REFERENCES

1. Depression, W. H. O. (2017). Other common mental disorders: global health estimates. Geneva: World Health Organization, 24(1).
2. Calvo, R., Milne, D., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685. https://doi.org/10.1017/S1351324916000383
3. Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and systems magazine, 6(3), 21-45. https://doi.org/10.1109/MCAS.2006.1688199
4. Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press. https://doi.org/10.1201/b12207
5. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1
6. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.
7. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In Proceedings of the international AAAI conference on web and social media (Vol. 7, No. 1, pp. 128-137).
8. Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, 115(44), 11203-11208. https://doi.org/10.1073/pnas.1802331115
9. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery, 8(4), e1249. https://doi.org/10.1002/widm.1249
10. Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering, 32(3), 588-601. https://doi.org/10.1109/TKDE.2018.2885515
11. Sau, A., & Bhakta, I. (2017). Predicting anxiety and depression in elderly patients using machine learning technology. Healthcare Technology Letters, 4(6), 238-243. https://doi.org/10.1049/htl.2017.0066
12. Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018, June). Deep learning for depression detection of twitter users. In Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic (pp. 88-97). https://doi.org/10.18653/v1/W18-0609

13. Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, 182-197. https://doi.org/10.1016/j.eswa.2019.05.023

14. S., M. (2021). Depressive/Non-Depressive Tweets between Dec'19 to Dec'20. https://doi.org/10.21227/9phc-ya88

15. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

16. Rokach, L. (2010). Ensemble-based classifiers. Artificial intelligence review, 33, 1-39. https://doi.org/10.1007/s10462-009-9124-7

17. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

18. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

19. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232. https://doi.org/10.1214/aos/1013203451

20. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785