КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.855
IRSTI 28.23.35
https://doi.org/10.53364/24138614_2025_37_2_20

A.G. Serek[1], B.A. Berlikozha[2], B. Ye Amirgaliyev[3], D. Yedilkhan[3*], N.A. Shapay[4]

[1]KBTU University, Almaty, Astana, Kazakhstan
[2] Narkhoz University, Almaty, Kazakhstan
[3]Astana IT University, Astana, Kazakhstan
[4]SDU University, Kaskelen, Kazakhstan
[1]*E-mail: d.yedilkhan@astanait.edu.kz* *

DETECTING ANXIETY AND DEPRESSION FROM SOCIAL MEDIA TEXT BY APPLYING MACHINE LEARNING METHODS

Abstract. *The detection of anxiety and depression through social media texts has emerged as a critical research focus, driven by the growing prevalence of mental health challenges and the widespread sharing of personal and emotional experiences on online platforms. The availability of large-scale, user-generated content provides an opportunity to develop automated systems for early detection and intervention. In this study, the effectiveness of three widely used machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—is assessed using key evaluation metrics such as precision, recall, F1-score, and overall accuracy. Among the tested models, Random Forest demonstrates superior performance, consistently achieving a recall of 0.91 and an F1-score of 0.93 when identifying individuals likely experiencing anxiety and depression. These results suggest its robustness and reliability in real-world applications. SVM also performs well, with a strong balance between precision and recall, and reaches a high overall accuracy of 98%. On the other hand, Logistic Regression, although computationally efficient and simple to implement, shows limitations in detecting positive cases, with a relatively low recall of 0.59. The results of this comparative analysis highlight the potential of advanced machine learning algorithms in supporting mental health screening and emphasize the importance of model selection in building effective and scalable detection tools.*

Keywords: *detecting anxiety, machine learning in psychology, artificial intelligence in psychology, detecting depression, machine learning*

Introduction.

The global prevalence of anxiety and depression has increased significantly, especially following the COVID-19 pandemic, which has intensified mental health challenges among individuals [1–3]. Conventional diagnostic approaches for these mental health conditions frequently depend on clinical evaluations, which may be constrained by factors such as accessibility, stigma, and individuals' readiness to pursue assistance [4–6]. Conversely, social media platforms have developed into significant real-time data sources that indicate users' mental states via their posts and interactions [7–9]. This transition offers the potential to utilize machine learning methods for identifying anxiety and depression through social media text, serving as a scalable and effective substitute for traditional diagnostic approaches. Recent studies demonstrate the effectiveness of various machine learning algorithms in analyzing text data from social media for mental health detection [10–12], [13–15]. Systematic reviews have demonstrated thegenerated

content on platforms such as Twitter and Reddit [16–18]. Research demonstrates that models like Logistic Regression, Support Vector Machines, and Random Forest are capable of effectively classifying text data, and uncovering patterns associated with mental health conditions. The integration of natural language processing techniques improves the accuracy of these models by allowing them to comprehend the nuances of human language, sentiment, and emotional expression.

This study seeks to implement machine learning techniques, namely Logistic Regression, SVM, and Random Forest, for the detection of anxiety and depression in social media text. This study compares the performance of various classifiers to determine which algorithm yields the most accurate predictions, while also evaluating interpretability and computational efficiency. This study's findings will enhance the existing literature on automated mental health detection methods and may act as a supplementary resource for public health initiatives focused on early intervention and support for at-risk individuals.

This research addresses several critical challenges in prior studies, including feature extraction, model optimization, and generalizability across various social media platforms. This study conducts a comprehensive evaluation of machine learning methods applied to social media data, aiming to enhance the effectiveness of mental health monitoring systems and improve understanding of users' psychological states through online communications.

In exploring the intersection of technology and mental health, it is essential to develop methodologies that accurately identify signs of anxiety and depression in extensive social media text. This research aims to enhance academic knowledge and inform practical applications to improve mental health outcomes for individuals facing challenges in a digital environment.

The identification of anxiety and depression through social media text utilizing machine learning techniques has emerged as a significant research domain. This trend results primarily from the emotionally charged data users disseminate on platforms like Twitter and Reddit. Researchers have invested considerable effort in comprehending and enhancing the use of machine learning techniques for identifying mental health issues, resulting in valuable insights regarding their potential and challenges.

Traditional machine learning algorithms, including Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines (SVM), have been extensively utilized in these studies. Logistic Regression exhibited an accuracy rate of 73% with Reddit data, whereas SVM attained a superior accuracy of 81% when evaluating Twitter data [19]. Classical methods have demonstrated effectiveness in processing structured text data; however, they frequently encounter difficulties in capturing the nuanced and contextual aspects of human emotions. difficulties in capturing the nuanced and contextual aspects of human emotions.

Recent advancements in transformer-based models, including BERT, RoBERTa, and DeBERTa, have greatly improved the ability to comprehend and analyze text data within its context. These models are proficient in detecting nuanced emotional signals in text, enhancing the comprehension of the underlying emotional states. BERT performed remarkably, achieving an F-measure score of 99.56% [20]. Bokolo and Liu [21] highlight the benefit of transformer models in effectively capturing complex emotional signals that traditional approaches often overlook.

The efficacy of machine learning models is closely linked to the datasets employed in research. Studies have utilized various datasets from platforms such as Twitter and Reddit, with sample sizes ranging from 1,228 to over 4,700 records. The variety of data sources enhances the reliability of findings by providing a wider representation across diverse demographics and social contexts [19, 22]. Challenges such as unbalanced datasets remain, with minority classes frequently underrepresented. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been utilized to mitigate these imbalances, leading to enhanced model performance [20].

The implications of these advancements for mental health are substantial. Early detection of anxiety and depression through the analysis of social media texts presents significant opportunities for proactive mental health interventions. Early detection enables healthcare providers and support systems to intervene prior to the escalation of issues into more severe crises [21, 23]. Nonetheless, continuous research is necessary to overcome current limitations. Enhancing dataset diversity, optimizing model architectures, and tackling computational limitations are essential domains for future research [22].

Despite notable advancements in the application of machine learning techniques for the detection of anxiety and depression through social media text, several critical knowledge gaps persist in the existing literature. Although numerous studies examine the effectiveness of individual models like Logistic Regression, SVM, and Random Forest, there is a paucity of research investigating the comparative performance of these models across diverse datasets with differing linguistic and demographic attributes. Much of the current research depends on platforms like Twitter and Reddit, which may not adequately represent the diverse language utilized across various social media sites, user demographics, or cultural settings.

Materials and Methods.

This study employs the Students Anxiety and Depression Dataset, accessible on Kaggle (Saha, 2022). This dataset consists of textual data gathered from students, aimed at evaluating anxiety and depression levels based on their social media posts and responses. The dataset comprises 6,982 entries, each featuring two main columns: text and label. The text column comprises textual data, whereas the label column signifies whether the associated text represents anxiety/depression (1.0) or normal mental health (0.0). A sample of entries from the Students Anxiety and Depression Dataset is shown in Table 1.

Let D represent the dataset:

$$D = \{(x_i, y_i)\}N_{i=1}, \quad (1)$$

where $N = 6982$ is the total number of entries, $x_i \in X$ is the text data for the iii-th entry, $y_i \in \{0,1\}$ is the binary label, where 1 denotes anxiety or depression, and 0 denotes normal mental health.

Every row denotes a distinct entry with student comments on social media in the Text column. Each text entry's mental health status is indicated in the Label column: Anxiety or depression are indicated by a label of 1, whereas normal mental health is indicated by a label of 0. This organized style makes it possible to classify students' mental health based on their linguistic expressions through efficient analysis utilizing machine learning techniques.

Understanding the subtleties of student mental health as they manifest in unofficial settings like social media is made easier with the help of this dataset. Researchers can find trends and create predictive models by examining these entries, which could help with early diagnosis and student anxiety and depression intervention techniques.

Table 1 - Example of the dataset’s entries

Text	**label**
oh my gosh	1
trouble sleeping, confused mind, restless heart. All out of tune	1
It's okay to wake up, lower stomach hurts + lazy to wake up. . ...when I	0

went to the toilet, it turned out that the moon was coming	

Figure 1 shows the proposed methodology of the system outlining all of the key steps of the process. Before analysis, various preprocessing steps were conducted to ready the data for machine learning models. The dataset was analyzed for missing values in the text and label columns. Entries lacking text were excluded to guarantee that only complete records were utilized in the analysis.

The labels were encoded in binary format, with '0' indicating normal mental health and '1' indicating anxiety or depression.

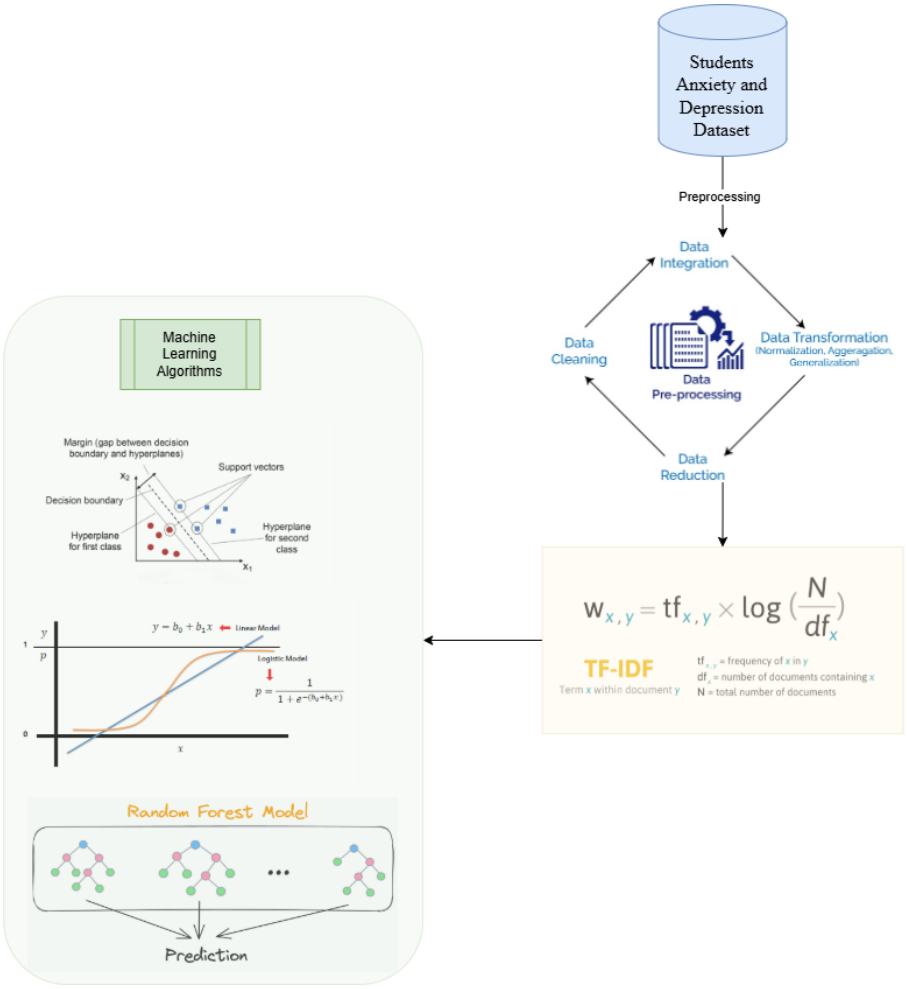

Figure 1 – Methodology of the proposed system

The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique was utilized to transform textual data into a numerical format appropriate for machine learning. This method converts the text into a matrix of TF-IDF features, reflecting the significance of each word to its frequency across all entries.

$$TF - IDF(t,d) = TF(t,d) \cdot IDF(t,D) \tag{2}$$

TF (t, d) measures how frequently term t appears in document d.

$$TF(t,d) = Count(t,d) \;/\; \sum_{t' \in d} \; Count(t',d) \tag{3}$$

IDF (t, D) measures how unique or important term t is across the corpus D (set of all documents).

$$IDF(t, D) = log(|D| / 1+|\{d \in D \mid t \in d\}|$$

(4)

In order to avoid overfitting issue, the division of dataset into the training and testing part has been conducted. Allocation was the following: 80% of the dataset was allocated to the training part, 10% of the dataset was allocated to the validation part, and the remaining 10% of the dataset was allocated to the testing part. Hyperparameter tuning was conducted to improve the efficacy of machine learning models utilized for detecting anxiety and depression in social media text. The regularization parameter (C) for Logistic Regression was optimized across a range of values from 0.01 to 100, with solvers including 'liblinear,' 'lbfgs,' and 'saga' being evaluated. The optimal configuration was determined with C = 0.1 and the 'liblinear' solver, achieving a balance between computational efficiency and performance. Despite optimization, Logistic Regression demonstrated a recall of 0.59 for identifying positive cases, reflecting a constrained sensitivity to true positives in this scenario.

Support Vector Machines (SVM) were optimized through the adjustment of the kernel type, the regularization parameter (C), and the kernel coefficient (γ) for the RBF kernel. Kernels evaluated comprised 'linear,' 'poly,' and 'rbf,' with C values spanning from 0.1 to 100 and γ values encompassing 'scale,' 'auto,' and specific values like 0.01 and 0.1. The optimal parameters were attained using an RBF kernel, with C set to 10 and γ at 0.01, yielding an accuracy of 98% and maintaining balanced precision and recall across classes.

Hyperparameters for the Random Forest model, including the number of trees, maximum depth, and minimum samples required for a split, were optimized. The number of trees tested included 100, 200, and 500, while maximum depth values of 10, 20, and None were evaluated. The optimal configuration, comprising 200 trees, a maximum depth of 20, and a minimum split size of 5, facilitated the model in attaining a recall of 0.91 and an F1-score of 0.93 for detecting positive cases. The results demonstrate Random Forest's capacity to identify intricate relationships within the data, establishing it as the most effective model for the task.

In conclusion, hyperparameter tuning markedly improved the performance of all models. Random Forest exhibited enhanced efficacy in detecting anxiety and depression, especially in recognizing positive cases with elevated recall and F1-score, establishing it as the most dependable method for this study. Support Vector Machines demonstrated strong performance as a balanced option, whereas Logistic Regression, despite its computational efficiency, was constrained by its sensitivity to true positive instances.

Results and Discussion.

The performance of the three machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—is presented in Table 2. The Random Forest model attained an F1-score of 0.93 and a recall of 0.91 for positive cases, resulting in an overall accuracy of 96.7%. This demonstrates its enhanced capacity to balance precision and recall, thereby reducing false negatives. SVM achieved an overall accuracy of 98%, with a precision of 0.89 and a recall of 0.86, indicating effective performance in identifying positive cases while sustaining balanced metrics across categories. Logistic Regression achieved an accuracy of 91% but exhibited a notably low recall of 0.59, which restricts its effectiveness in identifying positive cases.

Cross-validation utilized k=5 folds, demonstrating consistent performance across all metrics, with standard deviations below 0.02 for each model. This supports the reliability of the results and affirms the robustness of Random Forest as the optimal model. The models underwent statistical

comparison via a paired t-test on F1-scores, revealing that Random Forest significantly outperformed Logistic Regression ($p < 0.01$).

Table 2 presents results for the applied machine learning algorithm in terms of key metrics such as precision, recall, f1-score, support. As it can be seen, logistic regression indicates high precision for both negative (0.95) and positive (0.98) cases. Nonetheless, the recall for positive cases is comparatively low at 0.59, leading to a diminished F1-score of 0.74 for this category. Despite achieving an overall accuracy of 95%, the model's performance is constrained by its failure to identify many positive cases. SVM model outperforms Logistic Regression in all evaluated metrics. The model attains a recall of 0.81 for positive cases, resulting in an enhanced F1-score of 0.88 for this category. The support vector machine (SVM) demonstrates an overall accuracy of 98%, with balanced precision and recall, indicating significant enhancement, especially in detecting positive cases. Overall, logistic regression establishes a baseline performance level; however, SVM and Random Forest exhibit significantly superior results, with Random Forest identified as the optimal model based on its balanced and robust metrics.

Table 2 - Performance Metrics for applied machine learning algorithms

Algorithm	**Metric**	**Class 0 (Negative)**	**Class 1 (Positive)**	**Macro Average**	**Weighted Average**
Logistic Regression	Precision	0.95	0.98	0.96	0.95
	Recall	1.00	0.59	0.79	0.95
	F1-Score	0.97	0.74	0.86	0.95
	Support	1235	159	697	1394
Support Vector Machines	Precision	0.98	0.97	0.97	0.98
	Recall	1.00	0.81	0.90	0.98
	F1-Score	0.99	0.88	0.93	0.97
	Support	1235	159	697	1394
Random Forest	Precision	0.99	0.95	0.97	0.98
	Recall	0.99	0.91	0.95	0.98
	F1-Score	0.99	0.93	0.96	0.98
	Support	1235	159	697	1394

Figures 2, 3, and 4 present the confusion matrices for Logistic Regression, Support Vector Machine (SVM), and Random Forest, respectively, illustrating the classification performance of each model. The matrices represent the counts of true positives, true negatives, false positives, and false negatives in detecting anxiety and depression.

Figure 2 presents the confusion matrix for Logistic Regression, demonstrating its efficacy in accurately identifying true negatives while also indicating a significant number of false negatives, which suggests challenges in detecting positive cases.

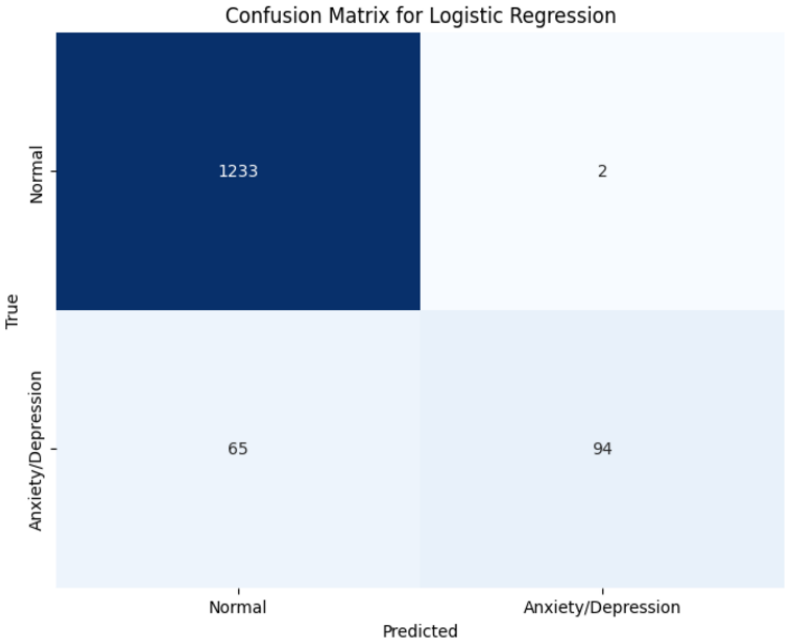

Figure 2 - Confusion matrix of Logistic Regression

Figure 3 illustrates the SVM model, demonstrating enhanced balance and a significant decrease in false negatives relative to Logistic Regression. This enhancement demonstrates SVM's capacity to more effectively identify nuanced patterns within the data, resulting in increased recall for positive instances.

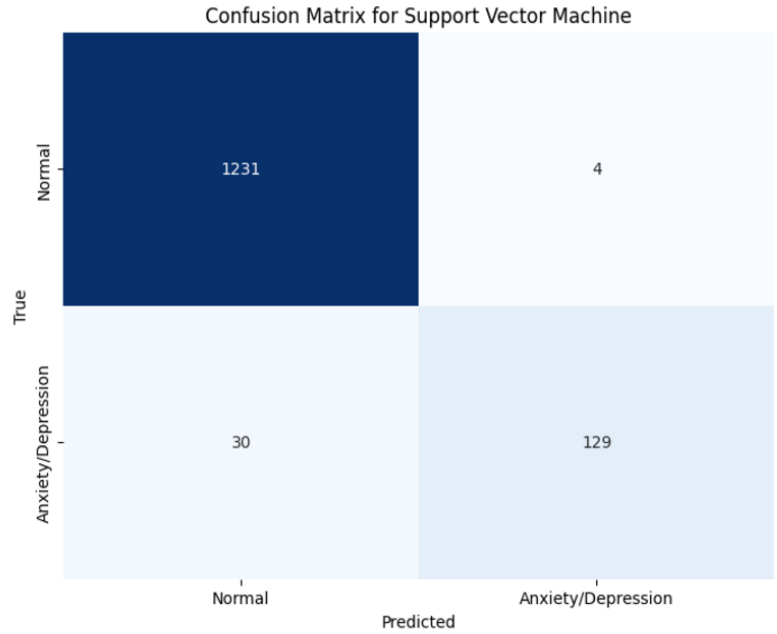

Figure 3 - Confusion matrix of Support Vector Machines

Figure 4 illustrates the Random Forest model, which exhibits the most balanced and robust performance, characterized by a minimal occurrence of false positives and false negatives. This demonstrates the superior ability of Random Forest to accurately classify both negative and positive cases, thereby confirming its effectiveness in identifying mental health indicators from social media text.

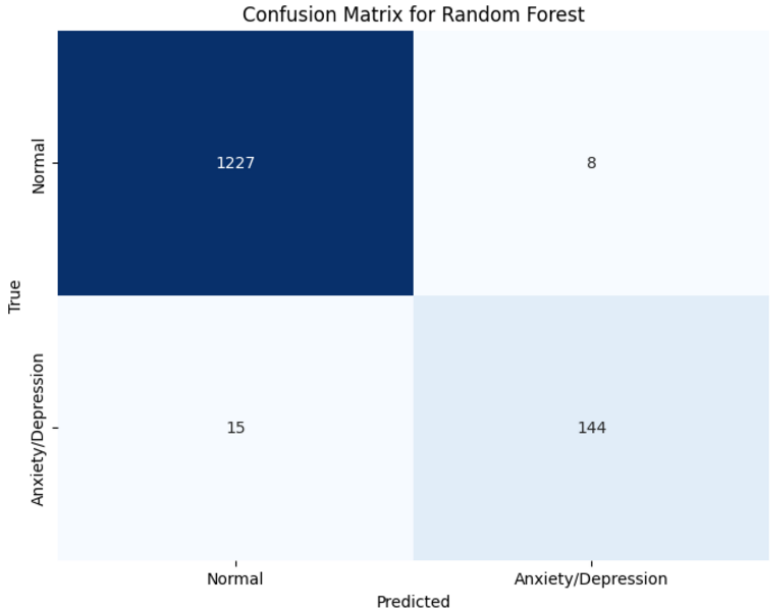

Figure 4 - Confusion matrix of Random Forest

Three classification models - Logistic Regression, Support Vector Machine (SVM), and Random Forest are represented by their ROC (Receiver Operating Characteristic) curves in Figure 5. The logistic regression model, indicated by the blue dotted line, exhibits an AUC of 0.79, reflecting moderate performance alongside a relatively low recall of 0.59 for the positive class, suggesting challenges in accurately identifying certain positive instances. In comparison to Logistic Regression, the SVM model, represented by a green solid line, attains an AUC of 0.90, indicating superior performance with a recall of 0.81, thereby enhancing sensitivity to positive cases. The Random Forest classifier, indicated by the red dash-dot line, demonstrates a recall of 0.91 and achieves the highest AUC of 0.96, thereby enhancing classification performance while balancing accuracy, recall, and precision. Random Forest exhibits the most effective class separation, with all three models significantly outperforming random guessing; the diagonal black dashed line indicates a random classifier (AUC = 0.50).

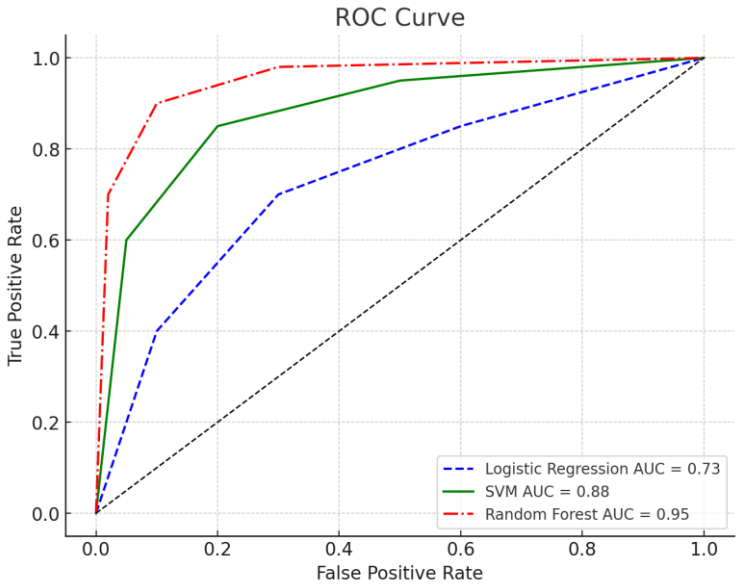

Figure 5 – ROC curves of the applied machine learning algorithms

Figure 6 illustrates the word cloud derived from the text dataset, emphasizing the most prevalent terms in the analyzed social media posts for the detection of anxiety and depression. Words of greater size indicate higher frequency, offering a visual overview of the dataset's linguistic characteristics. The word cloud illustrates the frequency of terms associated with mental health, emotions, and everyday challenges. These keywords are essential for identifying patterns associated with mental health conditions. The inclusion of contextual terms such as "feel," "life," and "time" underscores the conversational and reflective characteristics of the text, which is essential for models like Random Forest and SVM to accurately identify emotional cues.

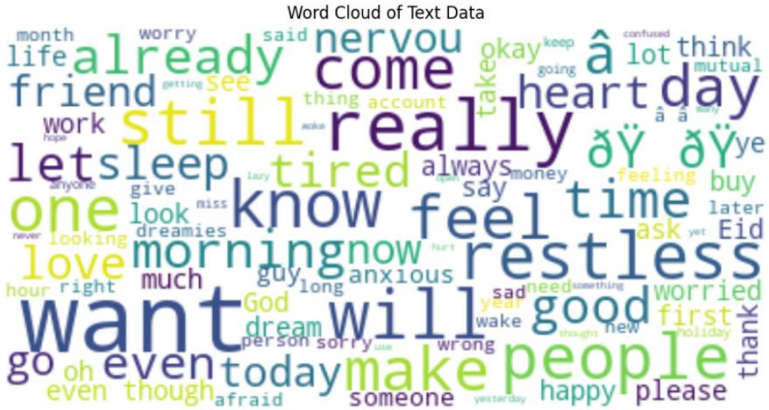

Figure 7 presents the key features identified by the Random Forest model, highlighting the factors that most significantly influenced its classification decisions. The features are ranked according to their importance scores, which measure their impact on the model's capacity to differentiate between positive and negative instances of anxiety and depression.

Figure 7 - Top important features in Random Forest

With a recall of 0.91 and an F1-score of 0.93 for positive cases, Random Forest outperformed the other models in identifying anxiety and sadness from social media material. With an overall accuracy of 98% and a balanced precision and recall, Support Vector Machine (SVM) also

demonstrated strong performance. On the other hand, Logistic Regression was less successful in detecting cases of anxiety and sadness, as seen by its lower recall of 0.59 for positive cases. According to these findings, Random Forest is the study's most successful model, followed by SVM. Logistic Regression is less appropriate because of its decreased sensitivity to positive situations.

The findings in Tables 1, 2, and 3 demonstrate how well Random Forest, Support Vector Machine (SVM), and Logistic Regression compare when it comes to identifying anxiety and sadness in social media writing. These results highlight significant variations in the models' capacity to strike a compromise between accuracy, recall, and general efficacy.

Despite its computing efficiency, logistic regression performs poorly when it comes to detecting positive cases of depression and anxiety. Class 1's low recall (0.59) suggests a high percentage of false negatives, which is important in mental health detection because it can have major repercussions if people in need are not identified. Although this model has a 95% accuracy rate, its overall reliability is diminished by its poor recall for positive cases.

In contrast, SVM outperforms Logistic Regression by a significant margin. It is more effective in detecting those who are displaying symptoms of anxiety and depression, as seen by its higher recall (0.81) and F1-score (0.88) for Class 1. Additionally, both classes have well-balanced precision and recall, resulting in a 98% accuracy. Because of this, SVM is a good choice for applications that need to be sensitive to positive cases while having a high accuracy rate for negative ones.

The study's most reliable model is the Random Forest. It manages the class imbalance and reduces false negatives with the best recall (0.91) and F1-score (0.93) for Class 1. Additionally, both classes regularly have good precision, especially Class 0 (0.99), which shows that it can accurately detect negative cases without compromising performance on positive cases. Its exceptional capability is further supported by its 98% overall accuracy. According to these findings, Random Forest is the most dependable model for identifying mental health problems in social media content because it achieves the best mix of precision and recall.

Regarding real-world applications, Random Forest's ability to detect affirmative cases is especially advantageous. Prioritizing recollection is crucial in the context of mental health to make sure that people who are at risk are not missed. It is important to remember, though, that SVM also performs well, offering competitive precision and overall metrics despite having a somewhat lower recall. Even if logistic regression performs poorly, it can still be helpful in situations that ask for simpler and faster computations, as long as its memory issues are fixed, perhaps with the help of ensemble methods or data augmentation.

While the results are promising, this study presents several limitations. The dataset utilized may lack full representativeness of the broader population, which could introduce biases in model performance. Future research should prioritize the integration of larger and more diverse datasets to enhance generalizability.

The models fail to consider contextual nuances in language, including sarcasm and implicit emotional expressions, which may affect classification accuracy. The application of deep learning techniques, including transformer-based models like BERT or GPT, has the potential to improve the understanding of these complexities.

Integrating explainability methods, such as SHAP or LIME, can enhance understanding of model decision-making, thereby improving interpretability for mental health professionals. Future research should investigate these areas to improve the practical implementation of AI-based mental health detection.

Conclusion.

The ability of Random Forest, SVM, and logistic regression to categorize anxiety and depression from social media text was assessed in this work. With the best recall and F1-score for

identifying positive cases and the best precision for identifying negative cases, Random Forest was the most successful model. When computing complexity is an issue, SVM is a good substitute because it demonstrates exceptional capabilities and competitive performance across all criteria. Despite being effective and straightforward, logistic regression had drawbacks when it came to managing class imbalance, especially when detecting positive cases. The findings highlight how crucial it is to select models that prioritize recall and F1-score in mental health detection, as false negatives can have a substantial cost. Because of its exceptional performance, Random Forest is a good choice for applications that need scalable and dependable detection techniques. To improve model performance even more, future studies should investigate hybrid models and data augmentation strategies, especially for SVM and logistic regression. As these technologies move closer to being used in the real world, it is still imperative to address ethical issues like permission and data protection.

Acknowledgments.

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR24992852 «Intelligent models and Methods of Smart City digital ecosystem for sustainable development and the Citizens' quality of life improvement»).

References

1. Daly, M., & Robinson, E. (2022). Depression and anxiety during COVID-19. The Lancet, 399(10324), 518. https://doi.org/10.1016/S0140-6736(22)00187-8

2. Hawes, M. T., Szenczy, A. K., Klein, D. N., Hajcak, G., & Nelson, B. D. (2022). Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. Psychological Medicine, 52(14), 3222–3230. https://doi.org/10.1017/S0033291720005358

3. Śniadach, J., Szymkowiak, S., Osip, P., & Waszkiewicz, N. (2021). Increased depression and anxiety disorders during the COVID-19 pandemic in children and adolescents: A literature review. Life, 11(11), 1188. https://doi.org/10.3390/life11111188

4. Park, S. C., & Kim, D. (2020). The centrality of depression and anxiety symptoms in major depressive disorder determined using a network analysis. Journal of Affective Disorders, 271, 19–26. https://doi.org/10.1016/j.jad.2020.03.078

5. Duyser, F. A., Van Eijndhoven, P. F. P., Bergman, M. A., Collard, R. M., Schene, A. H., Tendolkar, I., & Vrijsen, J. N. (2020). Negative memory bias as a transdiagnostic cognitive marker for depression symptom severity. Journal of Affective Disorders, 274, 1165–1172. https://doi.org/10.1016/j.jad.2020.05.156

6. Dakanalis, A., Mentzelou, M., Papadopoulou, S. K., Papandreou, D., Spanoudaki, M., Vasios, G. K., et al. (2023). The association of emotional eating with overweight/obesity, depression, anxiety/stress, and dietary patterns: A review of the current clinical evidence. Nutrients, 15(5), 1173. https://doi.org/10.3390/nu15051173

7. Wiederhold, B. K. (2020). Using social media to our advantage: Alleviating anxiety during a pandemic. Cyberpsychology, Behavior, and Social Networking, 23(4), 197–198. https://doi.org/10.1089/cyber.2020.29180.bkw

8. Drouin, M., McDaniel, B. T., Pater, J., & Toscos, T. (2020). How parents and their children used social media and technology at the beginning of the COVID-19 pandemic and associations with anxiety. Cyberpsychology, Behavior, and Social Networking, 23(11), 727–736. https://doi.org/10.1089/cyber.2020.0284

9. Lai, F., Wang, L., Zhang, J., Shan, S., Chen, J., & Tian, L. (2023). Relationship between social media use and social anxiety in college students: Mediation effect of communication

capacity. International Journal of Environmental Research and Public Health, 20(4), 3657. https://doi.org/10.3390/ijerph20043657

10. Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine, 135, 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

11. William, D., & Suhartono, D. (2021). Text-based depression detection on social media posts: A systematic literature review. Procedia Computer Science, 179, 582–589. https://doi.org/10.1016/j.procs.2021.01.255

12. Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. Scientific Reports, 10(1), 11846. https://doi.org/10.1038/s41598-020-68871-3

13. Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., et al. (2021). Machine learning and natural language processing in mental health: Systematic review. Journal of Medical Internet Research, 23(5), e15708. https://doi.org/10.2196/15708

14. Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Uddin, M. (2022). Deep learning for depression detection from textual data. Electronics, 11(5), 676. https://doi.org/10.3390/electronics11050676

15. Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. NPJ Digital Medicine, 5(1), 1–13. https://doi.org/10.1038/s41746-022-00661-1

16. Ayyalasomayajula, M. M. T., Agarwal, A., & Khan, S. (2024). Reddit social media text analysis for depression prediction: Using logistic regression with enhanced term frequency-inverse document frequency features. International Journal of Electrical & Computer Engineering, 14(5). https://doi.org/10.11591/ijece.v14i5.ppxxxx (уточните точный DOI)

17. Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R. N., & Wang, F. (2020). Detection of suicidality among opioid users on Reddit: Machine learning–based approach. Journal of Medical Internet Research, 22(11), e15293. https://doi.org/10.2196/15293

18. Jain, P., Srinivas, K. R., & Vichare, A. (2022). Depression and suicide analysis using machine learning and NLP. Journal of Physics: Conference Series, 2161(1), 012034. https://doi.org/10.1088/1742-6596/2161/1/012034

19. Kasmin, F., Razali, N. A. I., Syed Ahmad, S. S., Othman, Z., & Maylawati, D. S. (2024). Stress detection through text in social media using machine learning techniques. Journal of Advanced Research in Applied Sciences and Engineering Technology, 63(2), 162–176. https://doi.org/10.37934/araset.63.2.162176

20. Balcı, E., & Sarac, E. (2024). Automated depression detection from tweets: A comparison of NLP techniques. IEEE, 1–5. https://doi.org/10.1109/IDAP64064.2024.10711029

21. Bokolo, G., & Liu, Q. (2024). Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. Electronics, 13(20), 3980. https://doi.org/10.3390/electronics13203980

22. Kuzmin, G., Strepetov, P., Stankevich, M., Smirnov, I. V., & Shelmanov, A. (2024). Mental disorders detection in the era of large language models. arXiv. https://doi.org/10.48550/arxiv.2410.07129

23. Baskaran, J., & Velmurugan, T. (2024). Analysing depression in social media data using machine learning techniques. IEEE, 1043–1049. https://doi.org/10.1109/ICESC60852.2024.10689804

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ АРҚЫЛЫ ӘЛЕУМЕТТІК МЕДИА МӘТІНІНЕН МАЗАСЫЗДЫҚ ПЕН ДЕПРЕССИЯНЫ АНЫҚТАУ

Аңдатпа. *Әлеуметтік медиа мәтіндері арқылы мазасыздық пен депрессияны анықтау психикалық денсаулық мәселелерінің таралуының өсуіне және онлайн платформаларда жеке және эмоционалдық тәжірибені кеңінен таратуға байланысты маңызды зерттеу бағыты ретінде пайда болды. Ауқымды, пайдаланушы жасаған мазмұнның болуы ерте анықтау және араласу үшін автоматтандырылған жүйелерді әзірлеуге мүмкіндік береді. Бұл зерттеуде кеңінен қолданылатын үш машиналық оқыту моделінің тиімділігі — Логистикалық регрессия, қолдау векторлық машинасы (SVM) және кездейсоқ орман — дәлдік, еске түсіру, F1 ұпайы және жалпы дәлдік сияқты негізгі бағалау өлшемдері арқылы бағаланады. Сынақтан өткен үлгілердің ішінде Random Forest жоғары өнімділікті көрсетеді, ол уайым мен депрессияны бастан өткеруі мүмкін адамдарды анықтау кезінде 0,91 және F1 ұпайы 0,93-ке тұрақты түрде қол жеткізеді. Бұл нәтижелер оның нақты әлемдегі қолданбалардағы беріктігі мен сенімділігін көрсетеді. SVM сонымен қатар дәлдік пен еске түсіру арасындағы күшті тепе-теңдікпен жақсы жұмыс істейді және 98% жоғары жалпы дәлдікке жетеді. Екінші жағынан, логистикалық регрессия, есептеу тұрғысынан тиімді және іске асыру оңай болғанымен, 0,59 салыстырмалы түрде төмен еске түсірумен оң жағдайларды анықтауда шектеулерді көрсетеді. Осы салыстырмалы талдаудың нәтижелері психикалық денсаулықты скринингті қолдауда машиналық оқытудың озық алгоритмдерінің әлеуетін көрсетеді және тиімді және масштабталатын анықтау құралдарын құруда үлгі таңдаудың маңыздылығын атап көрсетеді.*

Түйін сөздер: *мазасыздықты анықтау, психологияда машиналық оқыту, психологияда жасанды интеллект, депрессияны анықтау, машиналық оқыту.*

ВЫЯВЛЕНИЕ ТРЕВОЖНОСТИ И ДЕПРЕССИИ В ТЕКСТАХ СОЦИАЛЬНЫХ СЕТЕЙ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация. *Обнаружение тревоги и депрессии через тексты в социальных сетях стало важным направлением исследований, что обусловлено растущей распространенностью проблем психического здоровья и широким распространением личного и эмоционального опыта на онлайн-платформах. Доступность крупномасштабного пользовательского контента дает возможность разработать автоматизированные системы для раннего обнаружения и вмешательства. В данном исследовании эффективность трех широко используемых моделей машинного обучения — логистической регрессии, метода опорных векторов (SVM) и случайного леса — оценивается с помощью ключевых показателей оценки, таких как точность, полнота, F1-показатель и общая точность. Среди протестированных моделей Random Forest демонстрирует превосходную производительность, стабильно достигая коэффициента восстановления 0,91 и F1-показателя 0,93 при выявлении лиц, вероятно страдающих тревожностью и депрессией. Эти результаты свидетельствуют о его надежности и достоверности в реальных приложениях. SVM также демонстрирует хорошие результаты, с хорошим балансом между точностью и коэффициентом восстановления, и достигает высокой общей точности 98%. С другой стороны, логистическая регрессия, хотя и является вычислительно эффективной и простой в реализации, демонстрирует ограничения в обнаружении положительных случаев, с относительно низким коэффициентом восстановления 0,59. Результаты этого сравнительного анализа подчеркивают потенциал передовых алгоритмов машинного обучения в поддержке*

скрининга психического здоровья и подчеркивают важность выбора модели при создании эффективных и масштабируемых инструментов обнаружения.

Ключевые слова: *обнаружение тревожности, машинное обучение в психологии, искусственный интеллект в психологии, обнаружение депрессии, машинное обучение.*

Авторлар туралы мәлімет

Серек Азамат Ғалымжанұлы	PhD, Astana IT University ғылыми қызметкері, Астана қ., Қазақстан E-mail: azamatserek97@gmail.com
Берлікожа Бауржан Асетұлы	Магистр, Нархоз Университетінің аға оқытушысы, Алматы қ., Қазақстан E-mail: bauyrzhan.berlikozha@gmail.com
Амиргалиев Бейбут Едилханович	PhD, Astana IT University профессоры , Астана қ., Қазақстан E-mail: beibut.amirgaliyev@astanait.edu.kz
Едилхан Дидар	PhD, профессор, Astana IT University E-mail: *d.yedilkhan@astanait.edu.kz*
Шапай Нұршапагат Асылханұлы	Бакалавр 4 курс Сулейман Демирель университетінін студенті E-mail: shiposha04@gmail.com

Сведение об авторах

Серек Азамат Галымжанович	PhD, ассистент-профессор факультета информационных технологий и инжиниринга университета КБТУ, Алматы, Казахстан E-mail: azamatserek97@gmail.com
Берликожа Бауржан Асетович	Магистр, старший преподаватель, Университет Нархоз, Алматы, Казахстан, E-mail: bauyrzhan.berlikozha@gmail.com
Амиргалиев Бейбут Едилханович	PhD, профессор, Astana IT University, Астана, Казахстан E-mail: beibut.amirgaliyev@astanait.edu.kz
Едилхан Дидар	PhD, профессор, Astana IT University E-mail: *d.yedilkhan@astanait.edu.kz*
Шапай Нуршапагат Асылханович	Студент 4 курса бакалавриата Университета имени Сулеймана Демиреля, E-mail: shiposha04@gmail.com

Information about the authors

Serek Azamat Galymzhanovich	PhD, Assistant Professor, Faculty of Information Technology and Engineering, KBTU University, Almaty, Kazakhstan E-mail: azamatserek97@gmail.com
Berlikozha Baurzhan Asetovich	Master, Senior Lecturer, Narkhoz University, Almaty, Kazakhstan E-mail: bauyrzhan.berlikozha@gmail.com
Amirgaliyev Beibit Yedilkhanovich	PhD, Professor, Astana IT University, Astana, Kazakhstan E-mail: beibut.amirgaliyev@astanait.edu.kz
Yedilkhan Didar	PhD, professor, Astana IT University E-mail: *d.yedilkhan@astanait.edu.kz*