INO TECH
Inspiring Technologies and Innovations

Inspiring Technologies and Innovations

June 2025, Volume: 4 Issue: 1

Research Article

Emotional State Analysis of Duzce University Students Using MODUM Application Data with Data Mining Methods

Diana REZEK[a], Oğuzhan KENDİRLİ[b]

[a]Duzce University, Graduate School of Natural and Applied Sciences, Department of Cyber Security, Düzce, TÜRKİYE

[b]Duzce University, Dr. Engin Pak Cumayeri Vocational School, Department of Electronics and Automation, Düzce, TÜRKİYE

ORCID[a]: 0009-0006-0844-3983
ORCID[b]: 0000-0001-7134-2196

Corresponding Author e-mail: diana97rezek@gmail.com

https://doi.org/10.5281/zenodo.15760668

Received	: 22.04.2025	Accepted	: 26.06.2025	Pages	: 25-32

ABSTRACT: The academic and social performance of students is significantly impacted by their mental health, which is a pivotal component of public health. It is well-documented that students in universities often undergo substantial psychological and emotional changes that can profoundly impact their daily behaviors and established mental health state. These alterations may manifest as a range of symptoms, including decreased focus and academic performance, as well as potential psychological discomfort and feelings of loneliness. The present study examines the psychological well-being of students at Duzce University in this regard. The study aims to categorize students' emotional states and identify those who could benefit from prompt psychological assistance. To this end, the study employs data mining and text mining techniques. This study was based on data previously collected from MODUM, an application used by Duzce University. To achieve this goal, the study integrates a range of data mining and text mining techniques for the classification of emotional states. The machine learning algorithms employed include Naive Bayes, Random Forest, Decision Tree (J48), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Logistic Regression. These algorithms were implemented across multiple software environments, such as Python, MATLAB, and R. Additionally, a variety of natural language processing techniques, including Bag of Words, TF-IDF, Word2Vec, and FastText, were used for effective text representation and preprocessing.

KEYWORDS: Data mining, sentiment analysis, text mining, classification.

1. INTRODUCTION

A growing body of research has documented an escalation in students' mental health challenges, attributable to factors such as peer pressure, academic pressure, and the transition to a novel study environment. These elements have had deleterious effects on students' mental and emotional well-being. In order to establish a secure and well-regulated learning environment, it is imperative to closely observe students' mental health and provide the appropriate level of support when necessary. A substantial proportion of programs are predicated on conventional therapeutic modalities that lack a foundation in behavioral analysis or real-time data. In this regard, the present study developed a sophisticated analytical model in which the psychological states of students may be tracked on a daily basis according to interaction with emojis and text communication analysis between students and the support team. The utilization of these algorithms is twofold: firstly, they assist in the classification of emotional states, and secondly, they facilitate the identification of cases necessitating proactive intervention. The objective of this project is to address the issue of "the lack of a smart and precise system to track and analyze the day-to-day psychological status of university students, and more unitarily identify vulnerable students. This predicament exemplifies a salient contemporary limitation of psychological support systems in academic institutions, which often resort to conventional methodologies or questionnaires that prove to be ineffective. The objective of this study is to propose a novel digital solution that higher learning institutions can utilize to make evidence-informed decisions using real-time data. Additionally, a range of sophisticated software programs, including R, MATLAB, and Python, were utilized to execute machine learning algorithms. A comparative intellectual approach was employed to analyze the processing of psychological and textual data, thereby establishing a novel methodological objective for the research. A review of the extant literature reveals that the majority of research in this area utilizes standard analysis tools or paper-based questionnaires, such as standardized questionnaire-based studies. The present study is distinctive in its integration of emoticon symbols and its utilization of the Turkish language, thereby ensuring a greater degree of alignment with actual reality. The present study proposes a practical model for the monitoring of the daily psychological state, the identification of potential data collected from an effective application to university (MODUM), the provision of an immediate categorization system for students requiring psychological support, and the offering of an analysis tool that will be available to Turkish universities for the university support system in the future [1].

The MODUM application was developed with the objective of providing students from all academic departments at Duzce University with assistance. Students are permitted to identify their emotional state by selecting an emoji a total of four times per day. The application provides a selection of nine emojis, and students have the option to request assistance and support from a dedicated team during standard university operating hours. The app's success underscores the importance of providing

psychological support to students to enhance their productivity and focus in academic life. The application integrates artificial intelligence and contemporary technological frameworks to optimize its efficacy [1].

Recently, deep learning models, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have achieved significant success in numerous domains due to their advanced capabilities in automatic feature representation learning. These models have also been extensively applied in various spatio-temporal data mining (STDM) tasks, such as predictive learning, anomaly detection, and classification. In this paper, we present a thorough examination of recent advancements in the application of deep learning techniques for STDM. The initial step involves the classification of the spatio-temporal data into five distinct categories. Subsequently, an overview of the prevalent deep learning models employed in STDM is provided. In the following section, an analysis of extant literature is conducted with the objective of identifying the types of spatio-temporal data, the data mining tasks, and the deep learning models [2].

Data mining is defined as the process of examining voluminous data sets with the aid of computer technology to discern patterns, trends, and insights. Data mining tools facilitate the prediction of future trends, enabling businesses to make informed, knowledge-driven decisions. The substantial volumes of data produced by conventional methods for predicting heart disease are too intricate and extensive to be efficiently processed and analyzed. The process of data mining provides the technological framework for the transformation of voluminous data sets into actionable information, thereby facilitating informed decision-making. The employment of data mining techniques has been demonstrated to expedite the process of predicting diseases, thereby enhancing the precision of the predictions. This paper presents a survey of studies that utilize data mining algorithms for the purpose of predicting heart disease. The employment of a data mining algorithm to predict future outcomes has been demonstrated to yield highly effective results. The application of data mining techniques to heart disease treatment data has the potential to yield results with a reliability comparable to that observed in prediction and diagnosis of heart disease [3].

This study provides a comprehensive review of state-of-the-art machine learning and data mining techniques used for medical diagnosis and prognosis, including neural networks, K-nearest neighbors (KNN), naïve Bayes, logistic regression, decision trees, and support vector machines (SVM).The findings of the present study demonstrate that neural networks consistently outperform other techniques in terms of diagnostic accuracy and predictive capacity, thereby demonstrating their robustness in handling high-dimensional and nonlinear medical data. This research underscores the potential of advanced machine learning algorithms in revolutionizing early diagnosis and effective prognosis, thus facilitating more personalized treatment plans and improved healthcare outcomes [4].

A pervasive issue that persists throughout students' academic trajectories is their substandard performance in high school. The ability to predict students' academic performance can benefit educational institutions in a variety of ways. Educational institutions can achieve their educational goals by providing support to students earlier in their academic careers. This support can be provided by identifying and understanding the factors that can affect the academic performance of students at the beginning of their academic careers. The objective of this study was to develop a model that could predict the achievement of early secondary students. Two sets of data were utilized for high school students who graduated from the Al-Baha region in the Kingdom of Saudi Arabia. In this study, three models were constructed using different algorithms: The following algorithms were utilized: Naïve Bayes (NB), Random Forest (RF), and J48. Furthermore, the Synthetic Minority Oversampling Technique (SMOTE) was employed to balance the data and extract features using the correlation coefficient. The performance of the prediction models has been validated using 10-fold cross-validation and direct partition, as well as various performance evaluation metrics, including accuracy curve, true positive (TP) rate, false positive (FP) rate, accuracy, recall, F-Measurement, and receiver operating characteristic (ROC) curve. The NB model demonstrated a prediction accuracy of 99.34%, closely followed by the RF model with 98.7% [5].

In this study, we compared several sampling techniques to address the varying ratios of the class imbalance problem (i.e., moderately or extremely imbalanced classifications) using the High School Longitudinal Study of 2009 dataset. To facilitate a comprehensive comparison, a multifaceted resampling approach was employed, encompassing random oversampling (ROS), random undersampling (RUS), and a synthesis of the synthetic minority oversampling technique for nominal and continuous data (SMOTE-NC) in conjunction with RUS, a hybrid resampling technique. The Random Forest was utilized as the classification algorithm to assess the outcomes of each sampling technique. The findings of the present study indicate that random oversampling for moderately imbalanced data and hybrid resampling for extremely imbalanced data appear to be the most effective approaches. The implications for educational data mining applications and suggestions for future research are discussed [6].

Sentiment analysis constitutes a pivotal component within the domain of natural language processing, encompassing the identification of a text's polarity, that is, the presence or absence of positive, negative, or neutral sentiment. The advent of social media and the Internet has led to a marked increase in the importance of sentiment analysis in a variety of fields, including marketing, politics, and customer service. Nevertheless, sentiment analysis poses significant challenges in the context of foreign languages, particularly in the absence of labeled data for training models. In this study, an ensemble model of transformers and a large language model (LLM) is proposed, with the model leveraging sentiment analysis of foreign languages by translating them into English. The languages employed in this study included Arabic, Chinese, French, and Italian, which were translated

using two neural machine translation models. LibreTranslate and Google Translate. Subsequently, an ensemble of pre-trained sentiment analysis models was employed to analyze the sentences for sentiment. Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, and GPT-3, which is a language model from OpenAI. The experimental results demonstrated that the accuracy of sentiment analysis on translated sentences exceeded 86% when employing the proposed model. This finding suggests that foreign language sentiment analysis is feasible through translation to English and that the proposed ensemble model outperforms independent pre-trained models and LLM [7].

The proliferation of unstructured data, manifesting as digitized text, is exhibiting a marked increase in terms of both volume and accessibility. Given the potential of text mining as a methodological framework, the primary objective of this manuscript is to empower novice and experienced innovation researchers to select, specify, document, and interpret text mining techniques in a manner that generates valid and reliable knowledge for the innovation management community. To this end, a systematic review of 124 journal articles was conducted, employing text mining techniques and published in a collection of 10 premier innovation management and 8 top general management journals. The results of the systematic manual and computational analysis of these articles illustrate the state and evolution of text mining applications in our field. They also allow for evidence-based recommendations regarding their future use. In this paper, we propose a set of methodological, conceptual, and contextual development priorities that we believe will contribute to establishing higher methodological standards in text mining and enhance the methodological richness in our field [8].

Text embedding models have been utilized in information retrieval applications, such as semantic search and question-answering systems based on retrieval-augmented generation (RAG). These models are typically Transformer models that have been fine-tuned with contrastive learning objectives. A particularly challenging aspect of fine-tuning embedding models pertains to the selection of high-quality hard-negative passages for contrastive learning. In this paper, we introduce a family of positive-aware mining methods that use the positive relevance score as an anchor for effective false negative removal, leading to faster training and more accurate retrieval models. An ablation study is conducted on hard-negative mining methods over their configurations, exploring different teacher and base models. We further demonstrate the efficacy of our proposed mining methods at scale with the NV-Retriever-v1 model, which achieves a score of 60.9 on the MTEB Retrieval (BEIR) benchmark and places first when it is published to the MTEB Retrieval on July 2024 [9].

Music has become an essential medium for the expression of emotions and the enhancement of human social experiences. However, the manual interpretation of emotions in song lyrics is often inaccurate and time-consuming, especially for complex or ambiguous lyrics. This creates a need for an automated system that can improve the accuracy and efficiency of emotion classification in song lyrics. Various algorithms, including K-Nearest Neighbor (K-NN), Naive Bayes Classifier, and Support Vector Machine (SVM), have been employed for the purpose of emotion classification in song lyrics. Preliminary studies have demonstrated that the integration of Support Vector Machine (SVM) with Particle Swarm Optimization (PSO) has been shown to attain an accuracy of up to 90%. In contrast, the application of K-Nearest Neighbor (K-NN) with feature selection has yielded the most optimal f-measure of 66.93%. Notably, the model exhibits superior performance in comparison to K-NN and Naive Bayes. The system implementation is web-based and utilizes the Streamlit framework, enabling users to input lyrics and obtain interactive emotion predictions. This research contributes to the analysis of music emotions and offers an efficient and more accessible alternative for emotion classification in song lyrics [10].

Digital transformation is a process that is causing rapid change around the world, especially in the development of metaverse technology. The advent of metaverse technology has elicited a mixed reception from the public, prompting a need for a thorough examination of public opinion regarding its acceptance or rejection. The objective of this research is to analyze 6,728 public comment data points regarding the metaverse on social media platform X, employing a text mining approach. The objective of this experiment is to identify the most effective text mining algorithm model for sentiment analysis in the metaverse. The findings will offer valuable insights to industry professionals engaged in metaverse development. Specifically, the precision value increased to 94%, the recall value increased to 93%, and the F1-score increased to 95%, as indicated by the confusion matrix. Conversely, the Naïve Bayes algorithm exhibited a comparatively lower accuracy of 91%, while the negative sentiment confusion matrix demonstrated an augmented precision of 87%, a heightened recall of 97%, and an augmented F1-Score of 92%. This enhancement in performance is indicative of the enhanced efficacy of the Naïve Bayes algorithm [11].

2. MATERIAL AND METHOD

The data for this research was provided by Duzce University in Turkey, with the understanding that all data will be kept confidential and secure in accordance with the university's privacy and data protection policies. A comprehensive dataset was collected from all faculties and academic levels. The dataset under study includes the number of emojis selected by students across all academic levels, including undergraduate, graduate, and diploma programs. Participants were able to express their emotional state on four separate occasions throughout the day. The university offers more than 55 undergraduate majors and over 50 associate degree programs across various fields, as well as several master's programs. This extensive selection of educational opportunities contributes to the university's extensive database, which is a valuable resource for researchers and students. The dataset under consideration comprised multiple principal columns, which were subsequently aggregated and examined, as illustrated in the following: The term "bolum" is used to denote a department or area of study.

Education_Level: The educational attainment of the subject is indicated by the following designations: The academic degrees offered at this institution include the Bachelor's, Associate's degrees, Master's, and Doctorate degree.

Each emoji is represented in a separate column, including: peace, anxiety, fear, happiness, anger, disgust, shame, sadness, and confusion. The column presents the number of times students in the major selected the emoji. For instance, a survey of 717 undergraduate Visual Communication Design students revealed a preference for the "anxiety" emoji. The "Total" column is used to indicate the total number of options in the "Emoji" section for each department. The " Positive_feeling_total " column was utilized to denote the aggregate number of positive emojis selected by students of the designated department.

The "Negative_feeling_total" column was utilized to denote the aggregate number of negative emojis selected by the department's students, which might encompass emotions such as sadness, anger, and fear. The mood_class column is a variable that serves to assess the prevailing psychological state of the department. The calculation of this index entails the aggregation of positive and negative feelings and their subsequent comparison. In the event that the negative feelings exceed the total positive feelings, the item is designated as "Negative." Conversely, if the negative feelings surpass the total positive feelings, the item is classified as "Positive."

A data set comprising text-based information from interactions between student-university psychological support team members was collected and subsequently analyzed. The objective of this analysis was to identify sentiment and patterns using text mining methodologies. A set of algorithms was utilized to assess the efficacy of classifying students' emotional states through the use of emojis. The objective of this analysis was to ascertain the most efficacious algorithm in terms of classification accuracy and the ability to support multiple datasets. The following algorithms were selected for inclusion in the study, and the scientific justification for each selection is provided below:

1. Logistic regression is a statistical model that has proven to be both simple and effective. This model is a valuable asset for comprehending the impact of a variable on the determination of a psychological state, offering a clear and readily explicable framework. This model has demonstrated a notable degree of efficacy in predicting binary taxonomic outcomes.
2. Support vector machine (SVM), The selection of SVM was predicated on its capacity to address intricate taxonomic challenges, whether linear or nonlinear, and its adeptness in managing textual data that had been converted into digital formats, such as TF-IDF and word2vec.
3. Random forest method, The proposed algorithm is a combinatory approach that addresses the challenges of overfitting and enhances precision by constructing a set of decision trees internally and then averaging their outputs to generate a predicted output. This approach produces an efficient and reliable method for classifying complex psychological states.
4. Naive Bayes algorithm: The selection of an efficient algorithm that operates with high speed and demonstrates particular efficacy in the analysis of text data is paramount. This is due to the proven effectiveness of the algorithm in text classification based on the fundamental principle of independence of variables.
5. The utilization of J48 decision trees is predicated on their capacity to explain the methods behind decision-making processes through the reiterated partitioning of data. This characteristic renders them more readily interpretable, making it easier to identify key classification variables.
6. Sixth, the focus was on the application of artificial neural networks (ANNs) in the context of deep learning. The objective of this examination was to assess the effectiveness of deep learning in analyzing nonlinear and complex relationships among data, particularly in multidimensional representations of text. This analysis draws parallels with the methods employed by Word2Vec and fastText.

Text mining methods: To analyze the text data from the student conversations with the psychological support team, a number of natural language processing (NLP) methods were employed.

The first method is the Bag of Words (BoW) approach, which is a text data representation technique that counts the number of occurrences of words without regard for their order.

The TF-IDF method is a linguistic analysis technique that facilitates the assessment of the importance of a word within a document by considering its relative frequency within that document in relation to the overall frequency of words.

The third method is Word2Vec, which involves representing words as numerical vectors. This method utilizes a neural network to simulate their semantic relations.

Fourthly, FastText is an extension of Word2Vec that facilitates the analysis of word subscripts, thereby conferring a performance advantage in rich conjugation languages, such as Turkish.

The following tools are used for programming and environment purposes:

The investigation utilized Python as the computer language of choice, employing the Anaconda environment to develop and execute data mining and text analysis algorithms through the utilization of Jupyter Notebook software. The R and MATLAB environments provided additional supplemental statistical and taxonomic analysis.

The performance of the algorithms was evaluated using precise metrics, which include:

a) Accuracy: is defined as the percentage of correctly classified cases out of all cases.

$$\textbf{Accuracy} = \frac{\mathbf{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad \textbf{(1)}$$

Where:
True positives (TPs) are cases that have been accurately categorized as positive.

True Negative (TN): The identification of negative cases that meet the established criteria

A "false positive" (FP) is defined as a positive case that has been incorrectly classified.

False Negative (FN): The term "negative situations" is employed in this text to denote circumstances that have been erroneously designated as such.

b) Precision: This calculation determines the percentage of positive cases that were accurately predicted out of all cases that were projected to be positive.

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\boldsymbol{TP + TN + FP + FN}} \quad \textbf{(2)}$$

c) Recall (Sensitivity): The objective of this analysis is to ascertain the extent to which the model is capable of accurately identifying positive cases. The following calculation method has been employed:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \quad \textbf{(3)}$$

d) F1 Score: A thorough evaluation of the model's classification performance was derived from the harmonic mean of precision and recall. The following equation is provided for reference:

$$\textbf{F1 Score} = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \quad \textbf{(4)}$$

The employment of a combined methodology enabled the study to provide an accurate and comprehensive evaluation of the psychological well-being of students at all educational levels. This objective was achieved by leveraging the complementary integration of text analytics and statistical data to enhance comprehension of emotional and psychological patterns. Consequently, this enhanced understanding informed psychological support strategies. The data were prepared using statistical analysis tools, employing the following techniques to implement classification algorithms:

Normalization: The objective is to standardize values.

Label Encoding: The objective is to transform categorical data into a numerical format.

The standard Gaussian Naive Bayes model was employed without the necessity for parameter adjustments, as it is a lightweight, fast, and efficient model, especially for text classification. The decision tree algorithm was executed with the following parameters: the splitting criterion was designated as 'entropy,' the maximum tree depth was set at 10, and the minimum number of samples required for node splitting was set at 2. These values were utilized to achieve an equilibrium between complexity and accuracy. In the random forest algorithm, the number of trees was set to 100, referred to as the "n_estimators" parameter.

The maximum depth, designated as max_depth, is set to 10. These parameters were modified to mitigate random bias and enhance the reliability of the results, particularly in the context of numeric data categorized by college and sentiment. In the Support Vector Machine (SVM) algorithm, the kernel type was identified as "linear" and "rbf." The gamma parameter is defined as "scale." In the Logistic Regression algorithm, the optimization method was set to "lbfgs," and the regularization value was determined as follows: The constant C is set to 1.0. This model was utilized in the context of binary classification (positive/negative) for the analysis of both numeric and textual data. In the Artificial Neural Networks (ANN) algorithm, the number of cells in the hidden layer was specified as: hidden_layer_sizes = (100).

The activation function is defined as "relu."

The training algorithm utilizes the "adam" solver.

The maximum number of iterations is set to 200.

The previous values were utilized within a rigorous and repeated training process, with the results being tested on a separate dataset to ensure the reliability of the models.

3. RESULTS

The study's findings encompass two primary components: the categorization of emoji-related statistical information and the sentiment analysis through text mining methodologies. The four primary performance metrics employed for the evaluation of each model were accuracy, precision, recall, and F1 score. The utilization of emoji statistical data for the purpose of classifying students' emotional conditions yielded disparate outcomes, contingent upon the employed algorithm and the designated programming platform. The study's foundation was a substantial dataset comprising student responses, utilizing nine discrete emojis to denote their emotional states. The dataset encompasses a diverse sample of students from various academic institutions and levels of study, providing a substantial and reliable depiction of the emotional landscape of the student body.

The logistic regression model exhibited a commendable capacity to predict cases, attaining a maximum accuracy of 78% and a maximum positive precision of 81%. The Support Vector Machine (SVM) and Random Forest machine learning algorithms

exhibited the highest recall rate of 100%, thereby demonstrating their capacity to detect all positive examples without missing any.

The logistic regression model exhibited an optimal balance between positive precision and recall, with an overall F1 value of 86%. The findings suggest that the logistic regression model attained balanced and superior overall performance.

Table 1. Comparison of algorithm performance.

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	78%	81%	78%	86%
Support Vector Machine	67%	67%	100%	79%
Random Forest	71%	70%	100%	81%

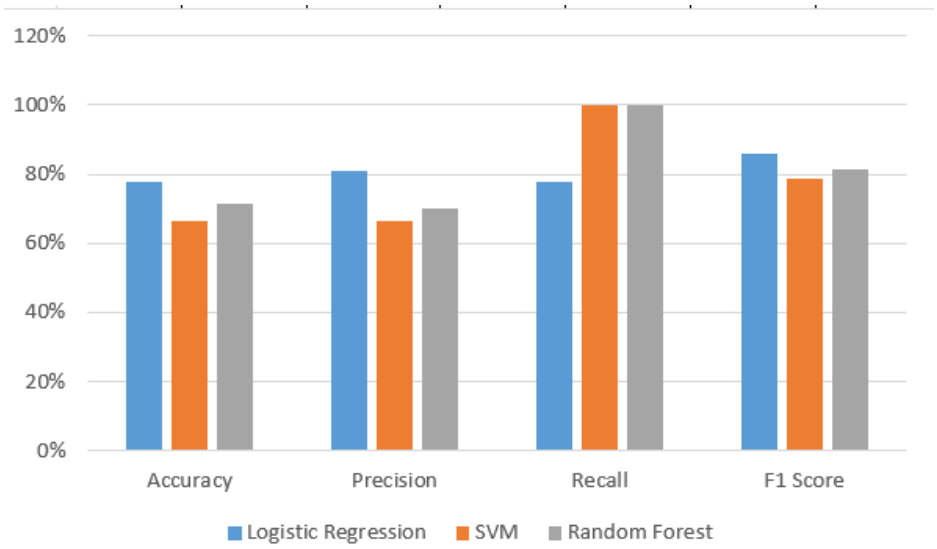

Figure 1. Comparing the performance of classification algorithms.

The SVM and random forest algorithms are particularly advantageous in scenarios where minimizing false negative errors is paramount. The application of sentiment analysis to text has been facilitated by the Python programming language, encompassing a range of text representation techniques, including both Bag of Words and FastText models, along with various classification algorithms. The findings of the study demonstrated that the integration of FastText representation and the Random Forest algorithm yielded the optimal classification accuracy of 100%. This outcome suggests that a state of perfect equilibrium was attained in the overall modeling performance. The combination of FastText and the decision tree algorithm (J48) also achieved the highest positive accuracy overall at 98%, indicating a very good capacity for correctly identifying positive cases. In the context of retrieval, the combination of bag of words with logistic regression emerged as a particularly salient approach, as it demonstrated the highest capacity for identifying relevant cases with exceptional efficiency.

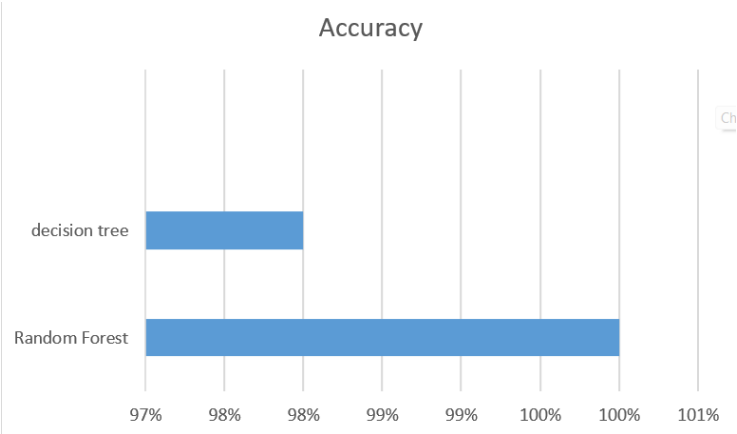

Figure 2. FastText text classifier performance.

The random forest model with bag of words also achieved the highest f1 scale, indicating its effectiveness in achieving a good equilibrium between positive accuracy and retrieval. As anticipated, the findings reveal that FastText representations, characterized by the lowest Root Mean Square Error, yield more precise predictions. However, the Bag of Words representation was found to be more successful in achieving a balance between retrieval and scales, which in turn produced improved accuracy. In the statistical data analysis portion of the study, the logistic regression model in MATLAB demonstrated the highest level of overall accuracy and the greatest F1 score. The SVM and random forest algorithms demonstrated a notable proficiency in looping, making them a particularly suitable option in scenarios where minimizing negative errors is of paramount importance. With regard to text analysis, the FastText representation demonstrated the highest classification efficiency, while the Bag of Words representation showed the greatest improvement in the F1 score and retrieval. This study is subject to several limitations related to the properties of the data utilized, which predominantly centered on emojis and abbreviated text. These factors may have an impact on the comprehensive emotional portrayal of students' experiences. The analysis exclusively incorporated aggregated data, precluding access to data that could have been more individual or chronologically sequenced. This limitation may have constrained our capacity to comprehensively understand the dynamic development of the psyche. The researcher posits that the scope of the study could be expanded in the future to encompass a more extensive array of behaviors. This expansion would facilitate the integration of sophisticated techniques from natural language processing, thereby enabling a more comprehensive and thorough analysis. Furthermore, the integration of multiple behavioral metrics could assist in more accurately identifying various psychological states.

4. CONCLUSION

This research has two important aspects that aim to explore the psychological states of Duzce University students using two different approaches. The first axis classifies the students' emotional states based on the statistical data of the emojis used by students within the mental health support application. The second axis categorizes emotions in conversational texts using text mining and natural language processing tools. The findings indicate that the quality of the input data and the programming environment can significantly impact the performance of categorization algorithms. Algorithms such as Random Forest and SVM demonstrate the highest recovery rates, while other models, including logistic regression, achieve the highest levels of accuracy and F1 score. In the textual emotional analysis, the FastText representation of the data proves to be notable, especially when combined with the Random Forest algorithm, which achieves accuracy rates of up to 100%.

A collegiate-level analysis of the findings shows that students demonstrating optimal psychological stability are predominantly enrolled in the faculties of Business Administration, Mathematics, International Trade, and Mechatronics Engineering. The underlying causes of this phenomenon appear to be multifaceted, including, but not limited to, the curriculum structure, the professional environment of the faculty, and the level of social and psychological support available within each faculty. It is important to underscore that the Departments of Political Science and Public Administration, Forest Engineering, Occupational Health and Safety, and International Relations display substantial indicators of psychological distress. Consequently, the respective faculty administrations, in coordination with mental health professionals, are advised to develop and implement targeted interventions aimed at promoting students' mental well-being. The findings highlight the significance of employing data and text mining methodologies in conjunction with artificial intelligence capabilities to promote mental well-being in academic settings. They emphasize that the caliber of outcomes and their practical application depend on the nature of the data collected and the path followed by the chosen analytical model. This highlights the necessity of identifying and applying scientific practices that align optimally with each study's objective and contextual needs.

In light of the findings and observations from this study, the following recommendations aim to enhance the mental well-being of students and optimize the efficacy of available supportive initiatives within higher education.

The initial step entails the provision of specialized psychological support to the institutions with the highest probability of being affected. The results of the study indicate that students specializing in Media, Forestry, Political Science, and Occupational Health and Safety exhibit a significantly higher prevalence of psychological distress compared to students pursuing other specializations. Therefore, it is recommended that departments within these faculties collaborate with the university's psychological counseling centers to provide individual consultations, ongoing support sessions, and interventions tailored to students' needs. The objective of this initiative is to expand the reach of the program to encompass all universities in Turkey. The program's success in improving student mental well-being at Duzce University suggests its potential applicability to other Turkish higher education institutions, provided that local administrative and cultural differences are carefully considered. This initiative facilitates the establishment of a comprehensive and sustained network of psychological support services. It is imperative to motivate students to proactively and effectively participate in the program. The efficacy of the platform depends on students' proactive engagement in regular assessments and reports regarding their psychological well-being. Consequently, it is essential to implement awareness initiatives that educate users about the platform's benefits and its commitment to privacy and data protection. These initiatives should also offer symbolic incentives to encourage regular engagement with the platform. The objective is to establish an environment characterized by amiability and inclusivity within the university setting. In order to alleviate the academic pressures experienced by students, it is imperative that institutions of higher education implement policies mindful of mental health. These policies encompass a range of initiatives, including providing flexible submission deadlines for assignments, incorporating stress and anxiety management courses into the curriculum, and integrating mental health services as a fundamental component of the university's daily operations, ensuring seamless and direct access for all students.

This study highlights the importance of using data and text mining techniques to effectively assess and support students' mental health. The study also highlights the importance of ongoing monitoring and collaboration between departments and psychological services to promote a healthy and supportive learning environment.

ACKNOWLEDGEMENT

I would like to express my profound gratitude to my eminent instructor, Dr. Oğuzhan KENDİRLİ, for his unwavering encouragement and invaluable contributions to this research study.

ETHICAL STANDARD DECLARATION

The necessary permissions were obtained from the Duzce University Ethics Committee to ensure the ethical use of the data collected for this study. The ethics committee approval, dated June 26, 2025, and numbered 588248, is available for review.

DATA AVAILABILITY STATEMENT

The data utilized in the study were obtained with the approval of the relevant ethics committee and in accordance with the principles of confidentiality. The dissemination of data is contingent upon adherence to Duzce University's data privacy and access policies.

REFERENCES

[1] D. Demirezen, Üniversite Öğrencilerinin Psikolojik İyilik Halini Belirlemek İçin Bir Mobil Uygulama Geliştirilmesi. Ph.D. Thesis, *Düzce Üniversitesi,* 2023.

[2] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. Article, *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 9, no. 2, pp. 447–459, 2018.

[3] K. K, M. M. Najumuddeen ve S. R, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. Makale, *International Journal of Data Mining Techniques and Applications,* Cilt 9, Sayı 1, ss. 250–255, 2020.

[4] M. Al-Batah, M. S. Alzboon, M. Alqaraleh, and F. A. Alzaghoul, Comparative Analysis of Advanced Data Mining Methods for Enhancing Medical Diagnosis and Prognosis. Article, *Data Metadata,* vol. 3, November 2024.

[5] A. S. Alghamdi and A. Rahman, Data Mining Approach to Predict Success of Secondary School Students: A Saudi Arabian Case Study. Article, *Educ. Sci.,* vol. 13, no. 3, 2023.

[6] T. Wongvorachan, S. He, and O. Bulut, A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Article, *Inf.,* vol. 14, no. 1, 2023.

[7] M. S. U. Miah, M. M. Kabir, T. Bin Sarwar, M. Safran, S. Alfarhood, and M. F. Mridha, A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Article, *Sci. Rep.,* vol. 14, no. 1, pp. 1–18, 2024.

[8] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. Article, *R D Manag.,* vol. 50, no. 3, pp. 329–351, 2020.

[9] G. D. S. P. Moreira, S. Paulo, B. Schifferer, and E. Oldridge, NV-Retriever: Improving text embedding models with effective hard-negative mining. Article, *Association for Computing Machinery,* vol. 1, no. 1.

[10] S. P. Rahayu, L. Afuan, G. A. Yunindar, E. Faculty, and U. J. Soedirman, "Implementation of text mining on song lyrics for song classification based on emotion using website-based logistic regression," *Jurnal Teknik Informatika,* vol. 6, no. 1, pp. 359–368, 2025.

[11] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse," *Jurnal Media Informatika Budidarma*, vol. 8, Apr. 2024, pp. 714–725.