Analyzing the Impact of Lifestyle and Psychological Factors on Student Depression using NLP Techniques

Abstract

In this competitive world, depression faced by the students become the most significant mental health challenge. The depression is caused by several factors such as academic pressure, lifestyle adaption, and psychological thoughts etc. This paper investigates the impact of lifestyle factors such as diet plans, exercise, body shaping, work balance and sleep patterns, and psychological factors such as suicidal thoughts, loneliness, genetic mental illness which leads to the risk of depression among the students. The paper utilizes a dataset comprising responses from 2,556 students which are a combination of structured and unstructured textual inputs for demographic, lifestyle and psychological in order to determine the risk factors involved in student depression. The paper further leverages NLP techniques to uncover the sentiments, dietary habits and sleep duration to understand the criticality of student depression risk and thus contribute to the development of mental health strategies for students.

Keywords: Student Depression, Lifestyle Factors, Psychological Factors, Natural Language Processing, Machine Learning

1 Introduction

Depression has become a serious health issue among students due to multiple challenges involved in academic, social and personal life. Their mental health is influenced by several factors like academic pressure, lifestyle and psychological strains. The World Health Organization has identified depression as a cause of disability which requires timely detection and cure. The paper tries to investigate into the relationships between several factors such as lifestyle factors (e.g. diet plan, sleep patterns etc.) and psychological factors (e.g. suicides and genetic mental illness etc.) which are associated with the risk of student depression. The patterns between different features of data are analysed by applying Natural Language Processing (NLP) techniques. NLP further helps to extract the meaningful patterns and sentiments involved in textual data in order to get insights into the emotional and psychological states of an individual. Further, by integrating NLP techniques with statistical and behavioural analysis helps to identify significant factors of depression risk and evaluate the extent to which different factors contribute to depression risk and further how their modifications can help in mitigating the risk. The paper contributes broadly to address student depression risk by providing insight into the impact of psychological and lifestyle factors, experimenting NLP techniques to research on student mental health with gaining deep insights into mental health policies.

The paper is divided into sections. Section 1 introduces the concept of student depression and how to combat the risk. Section 2 highlights the motivation for doing research. Section 3 briefs the literature review carried out related to student depression. Section 4 discusses the dataset utilized for determining the impact of student depression risk with several features included in the dataset. Section 5 briefs the research objectives carried out in the paper. Section 6 describes the methodology adopted to carry out the research objectives. Section 7 discusses in detail each research objective with their result outcome. Section 8 concludes the paper with the future research directions.

2 Motivation

Student depression has become a global concern for mental health issues. The student life which is affected by academic pressure, social expectations, lifestyle and dietary changes lead to psychological affliction. Despite the increase of awareness among students, detecting and preventing depression raises several challenges. Since the field of student depression is affecting various lives, there is a need to gain deeper insights into the approaches, factors affected, and how to combat depression among the students.

However, traditional statistical approaches provided insights but failed to determine emotions in textual data. This gap is addressed by introducing advanced Natural Language Processing (NLP) techniques, which helps to extract sentiments and opinions, investigate hidden patterns from textual inputs. This is utilized to analyse the impact of lifestyle and psychological factors such as suicidal thoughts, dietary plans, and sleep patterns which are connected with enhancing the mental well-bring.

The motivation of the paper arises from the need to:

1. Understand how different factors contribute to the emergence of student depression.
2. Extract and gain deep insights and patterns from textual data by utilizing advanced NLP techniques.
3. Determine useful findings related to preventive approach and target innovations related to student's mental depression.

The above objectives are addressed in this paper which further contribute to tackle mental health and student depression. The research emphasizes the importance of integrating computational tools and techniques with psychological insights to mitigate student depression risk.

3 Related Work

This section discusses some of the literature studies on detecting student depression. After reviewing the literature, it has been observed that previous studies have primarly focused on social media data and taking clinical transcripts.

In [1], Resnik et al. applied topic modeling approach to preprocess and analyze Reddit forums to determine the strong interdependencies between posts by user and symptoms of depression. Guntuku et al. [2] analyzed posts on Facebook posts to predict the risk of depression by analyzing the linguistic features. Bajaj et al. [3] classified level of depression by applying different machine learning techniques on the survey data collected from a significant number of users.

Other studies such as Al Hanai et al. [5] applied fusion of audio and textual features to detect depression. Yazdavar et al. [6] proposed a model based on semi-supervised learning to identify tweets related to depression. Shen et al. [8] investigated emotion detection through transformer-based architectures.

Dos Santos et al. [9] applied CNN-based sentence modeling to analyze the user sentiments. Tadesse et al. [10] surveyed different machine learning techniques that can be utilized for predicting mental health. Yourganov et al. [12] integrated the results of fMRI with machine learning techniques to predict depression. Lin et al. [13] employed techniques of deep learning on Weibo data to get an insights of mental health.

This paper tries to expand the existing work by integrating the structured survey data with their responses and analyze the risk of depression by applying techniques of NLP, sentiment analysis and topic modeling.

4 Description of the Dataset utilized

This section discusses the detailed study of the dataset utilized to address the student depression [15]. The dataset comprises around 2,556 student responses which are an amalgamation of demographic, lifestyle, and psychological data, as well as unstructured textual inputs. This dataset is utilized in order to investigate the importance of risk factors associated with student depression. The dataset comprises in total of 2556 entries, with 19 attributes which are an amalgamation of categorical, numerical, and textual data. The dataset is created from the student mental health survey as a part of research initiative to analyze risk factors contributed to student depression.

Different attributes of the dataset are discussed in detail below in Table 1.

Table 1: Brief Description of Attributes of Dataset

Category of Attributes	Attributes	Description on Attributes
1.Demographics	Age	Numerical data depicting the age of student
	Gender	Categorical data depicting the gender of student, e.g. male, female, transgenders etc.
	Educational Qualification	Depicts the educational qualification (undergraduate, postgraduate, doctoral or post-doctoral) and field area of research
2. Psychological Elements	Depression Status Indicator	It is binary indicator (yes/no) whuch depicts whether the student or participant is having depression or not.
	Suicidal Thoughts	It is a binary student response (yes or no) which indicates whether the student has experienced suicidal thoughts or not.
	Mental Illness History	It is a binary indicator (yes or no) whether there is a family history of mental illness involved or not
3. Lifestyle Factors	Dietary Habits	It is a categorical variable where diet type is categorized as “Healthy”, “Moderate” and “Unhealthy”
	Sleep Duration	It is a categorical variable which represents sleep duration which can be “less than 5 hours”, “5-6 hours”, “7-8 hours” or “more than 8 hours”.
	Frequency of Physical Activity	It describes the frequency of physical exercise done by an individual, which ranges from "Never" to "Daily"
4. Environmental and Academic Stress Factors	Academic Pressure	Student responses which provides details on the intensity of academic pressure factors
	Workload	It denotes the self report, depicting the number of work hours or study hours per week
5. Open Ended Responses		It decribes the experiences or emotional instances related to student's mental health

6. Missing Data		Dataset comprises some missing values which needs to be handled by imputation and removal methods

5 Research Objectives

This section briefs in detail the different research objectives carried out in the paper below:

RO1: Predicting the Depression Risk

This objective aims to identify different factors contributed to the risk of student depression and further predict the likelihood of depression. Available features such as **age, sleep duration, dietary habits, physical activity, etc.** are analyzed to predict whether the student is at the risk of depression or not. The objective is fulfilled by applying a **machine learning classification** approach which comprises algorithms including **data preprocessing**, **feature engineering**, **model training**, and **evaluation**.

RO2: Sentiment Analysis on Lifestyle Factors

This objective targets to perform sentiment analysis on lifestyle factors such as dietary habits, sleep duration etc. in order to assess their impact on student's depression. The objective is carried out by breaking the process down to several key steps which are detailed and analyzed in Section 7.

RO3: Topic Modeling on Open-Ended Responses

This objective tries to uncover hidden thematic structures in text corpora using topic modeling technique. The technique utilized for topic modelling on mental-health related textual data comprises open-ended responses to surveys or interview transcripts is **Latent Dirichlet Allocation (LDA).** The objective further allows us to identify groups of words that frequently co-occur in documents and to assign topics to each document.

6 Proposed Methodology

This section discusses the steps involved in the methodology utilized for carrying out the research objectives as illustrated in Figure 1.

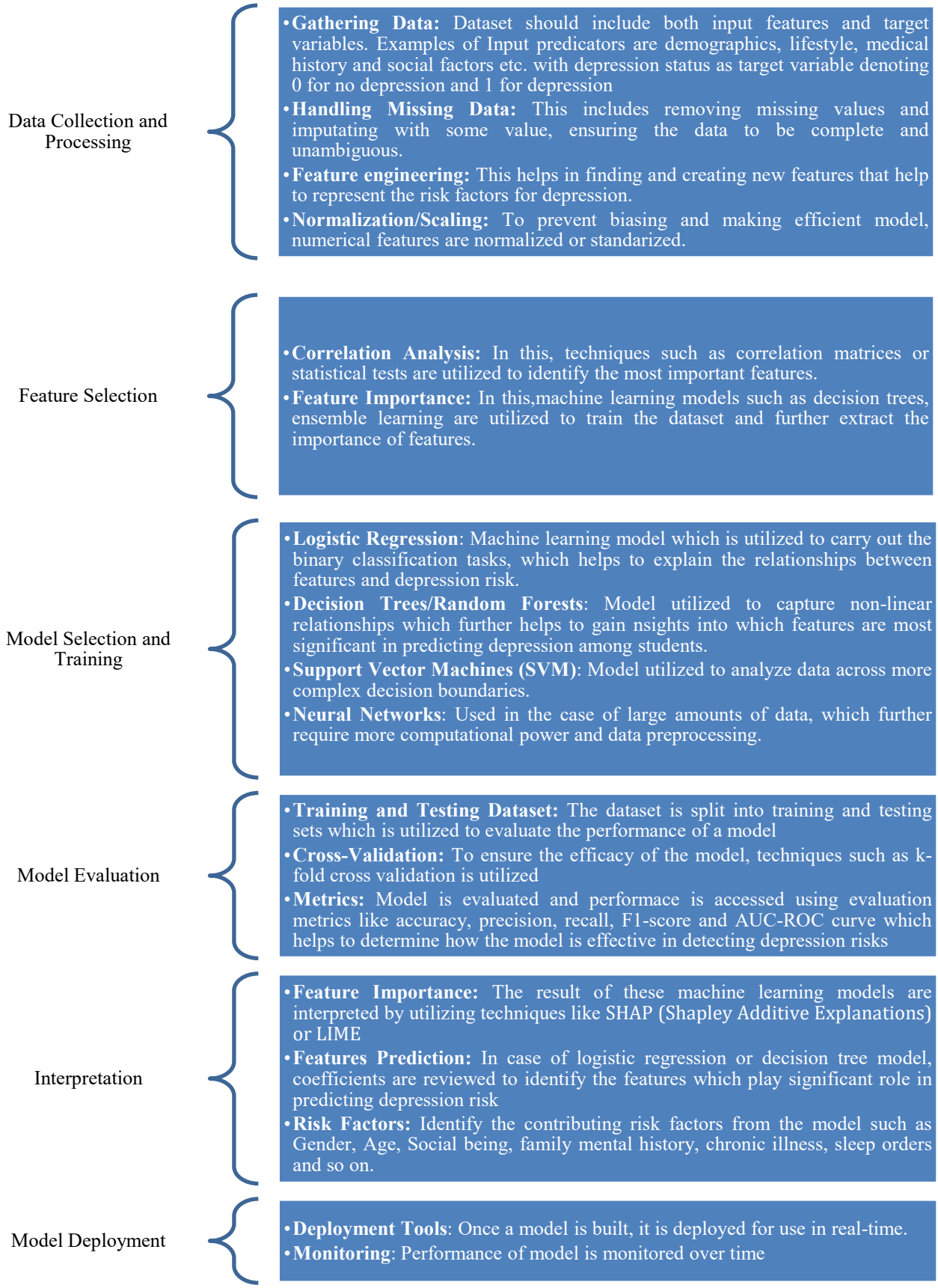

Figure 1: Methodology adopted to carry out the proposed research

7 Research Objective 1: Predicting the Depression Risk

This research objective focuses to identify key factors which contributes to the likelihood of depression among the students by applying machine learning classification techniques. This section discusses the steps which are used to implement an algorithm, such as data preprocessing, feature engineering, model training and evaluation. This objective targets to predict whether a student is at the risk of depression or not on the basis of features such as age, sleep, duration, dietary habits and physical activities and so on.

7.1. Step-by-Step Procedure

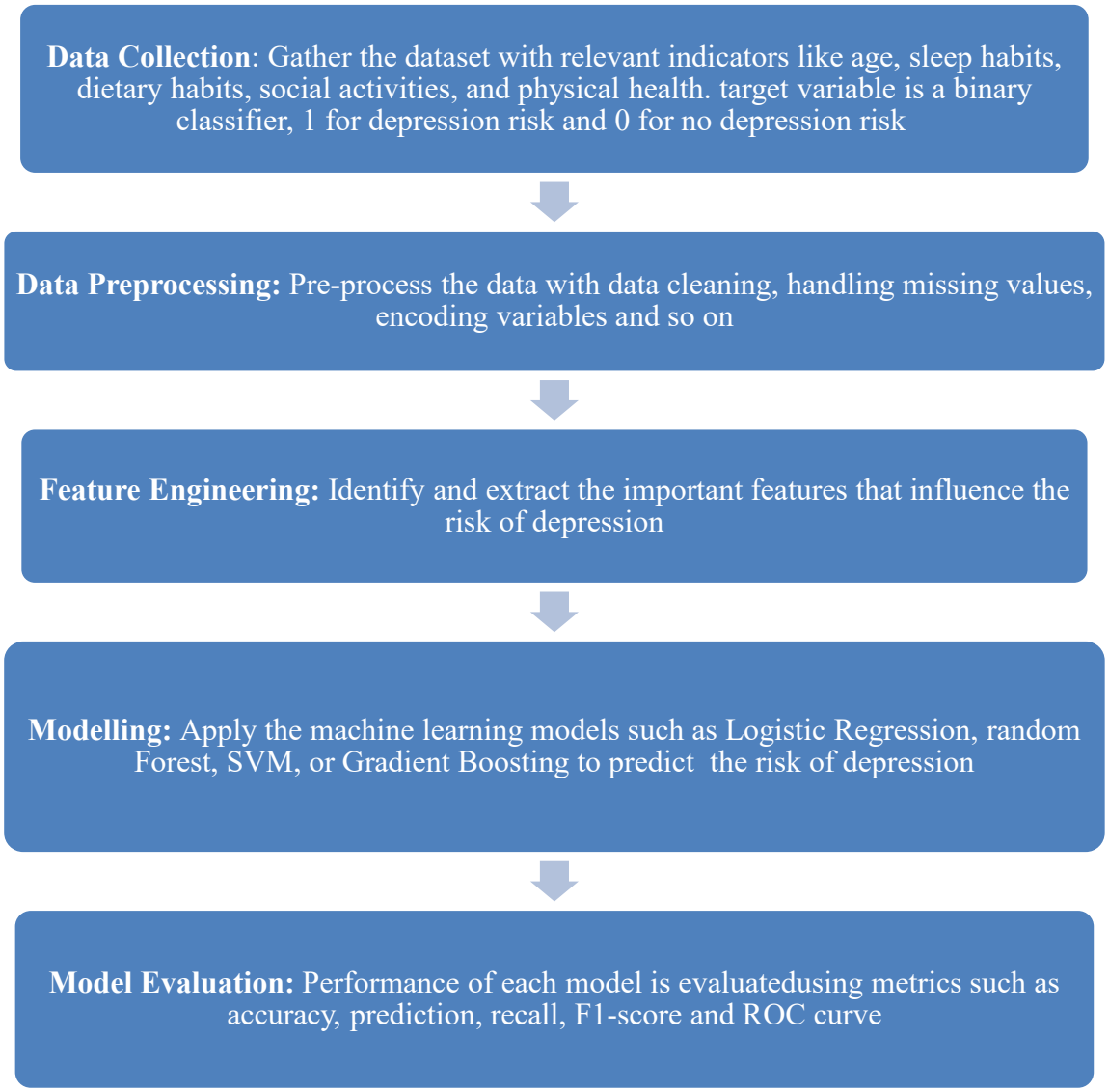

Figure 2: Step-by-step procedure to carry out the Research Objective 1

The above figure 2 depicts the step by step procedure utilized to predict the risk of depression among the students. The task is carried out a dataset downloaded from kaggle which comprises several components such as age, sleep habits, dietary habits, social activities, and physical health. The target is to depict whether there is a risk of depression or not, indicated by a binary variable, 1 for depression risk and 0 for no depression risk. The dataset is preprocessed where rows/columns are imputed with missing values, NULL values are replaced with mean value. After preprocessing, feature engineering is applied which depicts the important features which play significant role in determining the student depression risk. In this phase, the features which are not important are removed from the dataset in order to prepare the dataset for model evaluation. Finally, different machine learning models are applied and evaluated with the performance metrics that further helps us to determine which machine learning model play significant role in predicting the risk of depression among the students

7.2. Results Evaluation

This section discusses the performance of different machine learning models which are applied on the dataset to predict the student depression risk.

7.2.1. Support Vector Machine (SVM)

After applying SVM on the dataset, 88% accuracy is achieved with the detailing of confusion matrix and classification report

Confusion Matrix:

- True Negatives (TN) = 50: Correctly predicted no depression risk.
- False Positives (FP) = 5: Incorrectly predicted depression risk, but it was not
- False Negatives (FN) = 7: Incorrectly predicted no depression risk, but it was.
- True Positives (TP) = 38: Correctly predicted depression risk.

Classification Report:

Precision	Recall	F1-score	Accuracy
0.88	0.91	0.89	0.88
0.88	0.84	0.86	

7.2.2. Logistic Regression

Confusion Matrix:

- True Negatives (TN) = 48
- False Positives (FP) = 7
- False Negatives (FN) = 6
- True Positives (TP) = 39

Classification Report:

Precision	Recall	F1-score	Accuracy
0.89	0.87	0.88	0.87
0.85	0.87	0.86	

7.2.3. Random Forest Classifier (Ensemble Learning)

Confusion Matrix:

- True Negatives (TN) = 52
- False Positives (FP) = 3
- False Negatives (FN) = 4
- True Positives (TP) = 41

Classification Report:

Precision	Recall	F1-score	Accuracy
0.93	0.95	0.94	0.93
0.93	0.91	0.92	

7.2.4. k-Nearest Neighbors (KNN)

Confusion Matrix:

- True Negatives (TN) = 49
- False Positives (FP) = 6
- False Negatives (FN) = 8
- True Positives (TP) = 37

Classification Report:

Precision	Recall	F1-score	Accuracy
0.86	0.89	0.87	0.86
0.86	0.82	0.84	

7.2.5. Naive Bayes

Confusion Matrix:

- True Negatives (TN) = 50
- False Positives (FP) = 5
- False Negatives (FN) = 9
- True Positives (TP) = 36

Classification Report:

Precision	Recall	F1-score	Accuracy
0.85	0.91	0.88	0.86
0.88	0.80	0.84	

7.3 Interpretation of Results:

By the above results, it is clear that the machine learning techniques SVM and Random Forest performed the best if an accuracy is considered, by balancing the precision and recall. This states that the model is sufficient enough to correctly predicting the depression risk with fewer false positives and false negatives. In addition to this, Logistic Regression and Naïve Bayes perform well but having lower precision and recall for classes comprising students having depression risk. Lastly, KNN classification has lower accuracy but still it provides the reasonable balancing of precision and recall values. Thus, for predicted the student depression risk, it can be concluded that RandomForest appears to the most robust model, followed by SVM model. These results can further be improved by fine-tuning the hyperparameters for each model.

8. **Research Objective 2: Sentiment Analysis on Lifestyle Factors**

This research objective analyses the sentiments, emotions and polarity behind the responses which are related to lifestyle. These responses behind these lifestyle habits are identified from the keywords such as diet, sleep duration, frequency of exercise, calories intake etc. which influence understanding the impact on student depression.

8.1 Methodology

The sentiment analysis to understand the lifestyle habits is performed by applying NLP-based sentiment analysis models. The essential stages are:

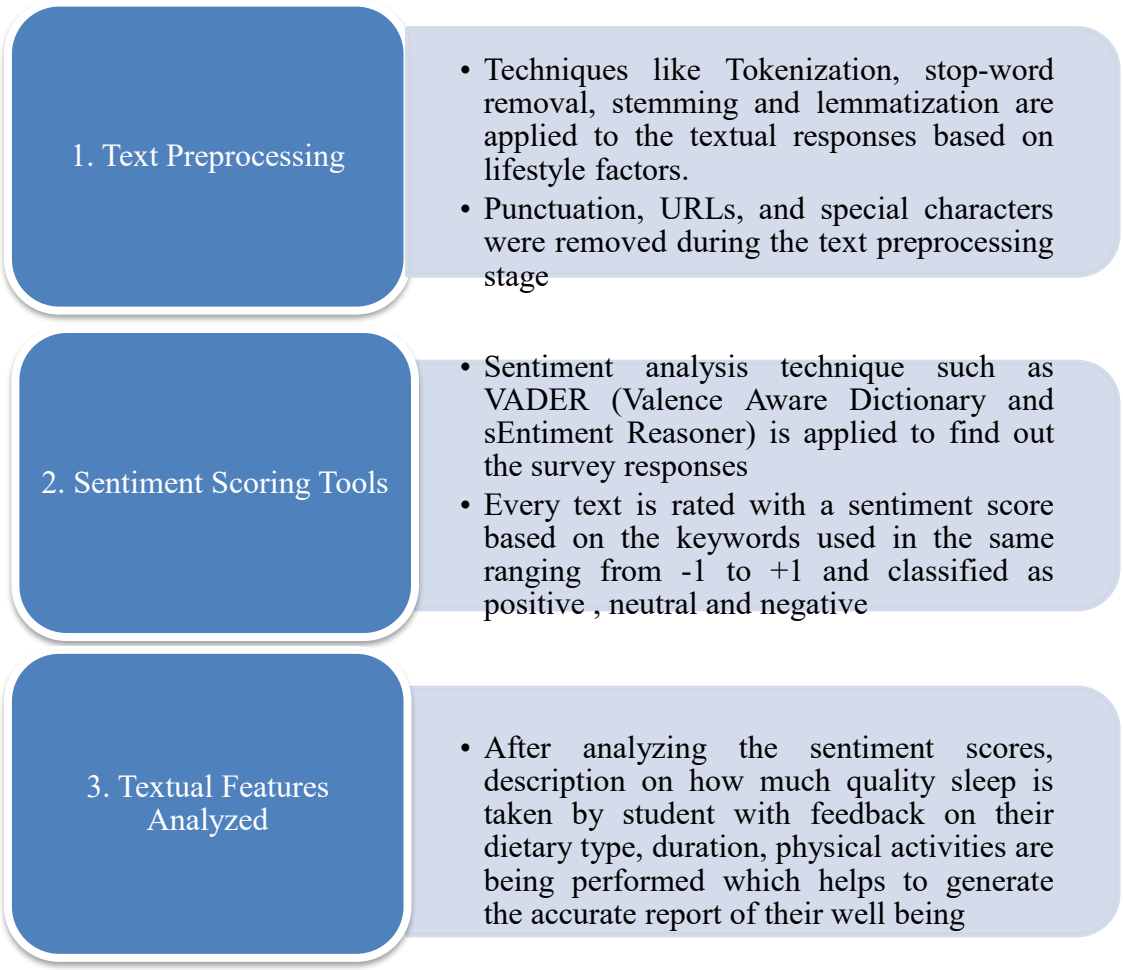

Figure 3: Stages to perform sentiment analysis on lifestyle factors

8.2. Results and Discussions

Sentiment Analysis helps to determine how the lifestyle choices correlate with the students' emotional well being. For example, negative sentiment scores depict the case of sleep deprivation, diet irregularity, minimal exercise, vitamin deficiency which acts as an early indication of the depression risk among students. Positive sentiment scores depict the case of healthy diet, sound sleep, exercise, good calorie intake. Thus task opens up a new indicator for designing sentiment-based alert systems to flag the depression risk among students.

The methodology applied to the dataset helps to attain the following results which are illustrated below in the table 2:

Table 2: Results of depression indicator with average sentiment score

Lifestyle Factors	Average Sentiment Score attained	Sentiment Type	Depression Indicator in %
1. Healthy Diet	+0.5	Positive	10%
2. Poor Sleep	-0.4	Negative	65%
3. No Physical Exercise	-0.35	Negative	58%
4. Regular Exercise	+0.3	Positive	15%

Thus, from the above results table, it is very much clear that the students who possess poor sleep and no physical activity possess negative sentiments and potentially increased chance of depression risk. Whereas, students who possess healthy diet intake and regular activities depict positive sentiments which further reflect the reduced chances of depression.

9. **Research Objective 3: Topic Modeling on Open-Ended Responses**

This research objective targets to determine the hidden concerns from the textual responses shared by the students. The terms related to lifestyle habits, dietary concerns, psychological and mental health struggles are identified to determine the related theme of concern. Thus, to extract such latent themes and topics from the responses, advanced techniques of topic modeling such as Latent Dirichlet Allocation (LDA) is employed.

9.1 Methodology

To carry out the task of topic modeling, the following procedure is adopted illustrated in Figure 4

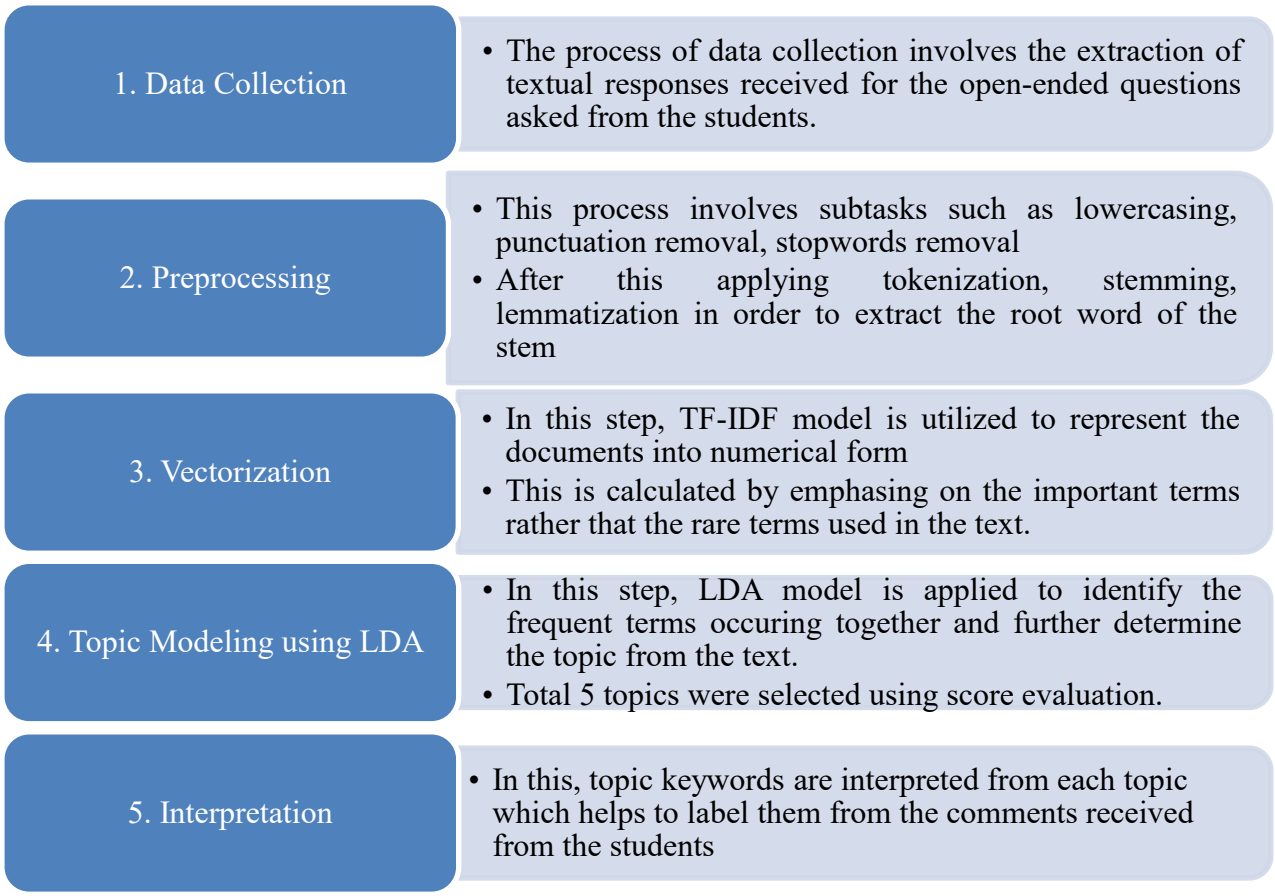

Figure 4: Procedure of topic modeling on text responses

9.2 Results of Interpretation

Table 3 below discusses the interpretation received from the topic keywords identified from each topic of interest which helps to identify the students concerns and further help health professionals to design support systems that can understand the student concerns from the textual data.

Table 3: Interpretation from topic keywords

Topic	Top Keywords	Interpretation
T1	anxiety, overwhelmed, sleep, stressed, exam	Academic pressure and anxiety
T2	therapy, counsellor, support, emotions	Seeking emotional support and therapy
T3	medication, side effects, dosage, doctor	Mental health treatment and medical care
T4	alone, isolated, friends, distance, social	Social isolation and loneliness

Topic	Top Keywords	Interpretation
T5	mindfulness, healthy, workout, meditation, routine	Lifestyle improvements and coping mechanisms

The above table concludes that the academic anxiety (T1) and social isolation (T4) became evident as dominant themes. Therapy and self-care (T2, T5) evolved as coping methods. In addition to this, some student expressed their concerns about the side-effects of medication they are undergoing, leading to the challenges of pharmacological concerns.

10. Conclusion and Future Research Directions

This paper utilizes the advanced techniques of Natural Language Processing and machine learning to analyze the factors which contribute to the risk of student depression. The methodology of structured data, evaluating sentiments and unstructured mining is utilized. The paper determines how computational linguistics approach helps to gain deep insights into the mental health challenges which are faced by the students. RO1 targets on predicting the risk of depression by applying machine learning models in which Random Forest and SVM outperform. RO2 analyses the indicator of depression by applying sentiment analysis to the textual response received corresponding to the open-ended questions asked by the students. RO3 determines the recurring topics by applying topic modelling techniques such as LDA.

Thus, these findings depict that the integration of computational models with psychological insights helps to build a robust framework which helps to early determine the target of student depression risks.

This research further opens various directions to gain valuable insights such as (1) the health data can be extracted from wearable devices (e.g., sleep trackers, heart rate monitors) to perform real-time monitoring which helps to assess the probability of depression risk and can be immediately intervenes, (2) Since the nature of mental health is dynamic, dynamic topic modelling techniques can be applied that helps to determine the trend of psychological patterns among the students, (3) dataset can be further expanded and multilingual tool can be designed that will enhance the applicability of the tool. (4) Recommendation system can be built on which helps to generate the individual insights and suggest medications based on each students' experience. Thus, this research can pave way for progression in the field of mental health and psychology.

Conflict-of-Interest:

- The author has no competing interests to declare that are relevant to the content of this article.
- The author certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

[1] P. Resnik et al., "Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter," in Proc. CLPsych, 2015.

[2] D. Guntuku et al., "Detecting depression and mental illness on social media: an integrative review," Curr. Opin. Behav. Sci., vol. 18, pp. 43–49, 2017.

[3] N. Bajaj et al., "Machine Learning Techniques for Depression Detection Using Survey Data," IJACSA, vol. 12, no. 4, pp. 234–241, 2021.

[4] X. Wang et al., "Using LSTM models for sentiment classification in healthcare domain," IEEE Access, vol. 8, pp. 132938–132946, 2020.

[5] M. Al Hanai et al., "Detecting Depression with Audio/Text Sequence Modeling of Interviews," Proc. Interspeech, 2018.

[6] M. Yazdavar et al., "Semi-supervised approach to model depression symptoms in Twitter users," IEEE Int. Conf. Big Data, 2017.
[7] G. Coppersmith et al., "Quantifying mental health signals in Twitter," ACL Workshop CLPsych, 2014.
[8] G. Shen et al., "Deep learning framework for identifying depressive users on Twitter," Proc. WWW Companion, 2017.
[9] C. dos Santos et al., "Classifying short texts using CNNs," arXiv preprint arXiv:1408.5882, 2014.
[10] M. M. Tadesse et al., "Detection of depression-related posts on social media using machine learning techniques: A survey," J. Depress. Anxiety, vol. 7, no. 1, pp. 1–9, 2018.
[11] S. Rude et al., "Language use of depressed and depression-vulnerable college students," Cogn. Emotion, vol. 18, no. 8, pp. 1121–1133, 2004.
[12] Y. Yourganov et al., "Machine learning with neuroimaging for diagnosis and treatment of depression," NeuroImage Clin., vol. 20, pp. 102–111, 2018.
[13] H. Lin et al., "User-level psychological stress detection from social media using deep neural network," Proc. ACM IMWUT, vol. 1, no. 3, pp. 1–30, 2017.
[14] A. Orabi et al., "Deep learning for depression detection of Twitter users," Proc. WorldCIST, 2018.
[15] Kaggle Dataset: Student Mental Health Survey, Available at: https://www.kaggle.com/datasets/souvikbra/student-mental-health
[16] C. Park et al., "Emotion Recognition and Analysis on Student Feedback using NLP and Deep Learning," IEEE Access, vol. 9, pp. 77589–77600, 2021.
[17] H. Sadeghi et al., "A Survey on Sentiment and Emotion Analysis for Depression Detection in Social Media," ACM Comput. Surv., vol. 55, no. 2, pp. 1–38, 2023.
[18] S. Bi et al., "Detecting early signs of depression from student writings using transfer learning," J. Affective Disorders, vol. 302, pp. 119–130, 2022.
[19] B. Settouti et al., "Hybrid models for predicting psychological disorders using NLP," Health Informatics J., vol. 28, no. 3, 2022.
[20] J. Lee et al., "Mental health monitoring system using wearable sensors and NLP," Sensors, vol. 23, no. 4, 2023.