A Machine Learning Method for Detecting Depression Among College Students

Peter J. Yu
Louisiana School for Math, Science, and the Arts
Natchitoches, Louisiana

ABSTRACT

As depression is becoming more prevalent on college campuses, it is increasingly a critical topic to investigate. Recently, studies using machine learning techniques have begun to predict depression and other mental illnesses. However, there is little understanding of why these mental problems occur. In this study, the causation of depression among college students posting on the popular social media platform Reddit is studied, and several machine learning classifiers for depression detection are compared. Of the 7,680 semi-anonymous Reddit posts examined, 552 contained depression-related keywords. After applying a series of natural language processing (NLP) techniques, three primary areas of depression were found among college students: institutions and programs; academic projects and assignments; and the college environment. Moreover, the results of this study show the effectiveness and performance of different machine learning classifiers. The classifier with the highest accuracy was Adaptive Boosting (AdaBoost), detecting depression with 99% accuracy, while the Random Forest classifier had the highest F1 score of 1.0.

Keywords

College, College Students, Depression, Mental Health, Machine Learning, Natural Language Processing (NLP), Latent Dirichlet Allocation (LDA), Social Media, Reddit

1. INTRODUCTION

Depression is a serious medical illness affecting how people feel, act, and think [1]. Since the COVID-19 pandemic, depression has become more prevalent among college students. Recent studies have found that 1 in 3 college students experience significant depression, and up to 44% of college students have symptoms of depression [2]. Unfortunately, although many college students struggle with depression and other related mental illnesses, they are reluctant to ask for help. As a result, the risks of harmful outcomes and detrimental behaviors are increased. In 2020, only an estimated 66% of U.S. adults aged 18 or older with mental illnesses received treatment for major depressive episodes [3]. Since depression impacts college life to a great extent, in-depth research is essential to address this issue.

As postsecondary education is commonly perceived to make an individual more successful, many students choose to leave their homes to obtain a college degree. Although the transition to college can be an exciting new experience, it can be intellectually and socially stressful [4]. All students will face the challenges of adjusting to new routines and establishing new relationships [5]. Acclimation to the college environment can produce negative consequences and induce homesickness among students. Homesickness has been linked to various psychological problems, including social anxiety, lack of concentration, and depression [6]. Adjusting to the college environment is, therefore, critical to investigate as it is associated with depression.

College students exposed to academic stressors such as time management, course load, classroom competition, and academic pressures cite education as a primary source of declining mental health [7]. However, in addition to academic factors, several other aspects of college life may increase the likelihood of depression and mental illnesses. For example, insufficient family and social support have been positively correlated with high levels of depression [8], as well as feelings of inferiority, low satisfaction in life, and dissatisfaction with body image [9]. Studies also indicate that poor diet and sleep habits have a negative impact on mental health and academic performance [8][10]. These factors and the lack of exercise and smoking are correlated with depression, and those suffering from depression have higher rates of cancer, infection, heart attacks, and mortality [11].

It is clear that understanding the issues related to depression is crucial for improving college students' psychological and physical well-being. The causation of depression is particularly important to investigate because colleges and other academic institutions can encourage behaviors that help prevent mental illnesses. This study aims to provide a better understanding of the factors that trigger depression in the college environment and introduces a new methodology for detecting depression among college students. Specifically, a series of natural language processing (NLP) techniques were used to identify depression-related areas in the posts by college students on Reddit, a popular social media platform. In addition, machine learning classifiers (e.g., Logistic Regression, Support Vector Machine, Random Forest, and Adaptive Boosting) were trained and tested on the dataset to detect depression. The results of this study show that the machine learning methodology is promising for exploring issues related to depression among college students and depression detection.

2. LITERATURE REVIEW

In recent years, depression among college students has gained more attention in academic research. There are various types of studies examining the areas in a college environment that are relevant to depression and other mental illnesses. For instance, Beiter et al. [4] investigated the prevalence and correlates of stress, anxiety, and depression in 374 undergraduate students aged 18 to 24. In order to identify the correlates of depression, a survey was conducted for the students to rate the level of concern for each challenge relevant to daily life and consisted of demographic and Depression Anxiety Stress Scale (DASS21) questions [4]. The results of Beiter et al. show that the top three concerns of college students were academic performance, pressure to succeed, and post-graduation plans [4]. Furthermore, the study found that upper-level students, transfers, and students living off-campus were the most depressed [4]. Although Beiter et al. identified the depression-related concerns of college students, the small sample used in the study was limited as it was collected from one institution.

Ebert et al. [12] assessed the incidence of major depressive order (MDD) among 2,519 first-year college students. A multivariate model was estimated with socio-demographic variables and depression-related experiences to predict MDD in new university students [12]. Ebert et al. established that social-demographic characteristics such as gender, age, and parent education did not have a notable impact on predicting MDD. Moreover, the study's results estimated the incidence of MDD at 6.9%, and of the 10% of students with the highest risk of MDD, around one out of four developed MDD [12]. Regarding the most significant factors, suicidal plans or attempts were most strongly associated with MDD. It is important to note that the response rates for Ebert et al. were modest (61.0% at baseline; 57.5% at follow-up), which resulted in bias [12].

With the development of social media and the internet, online networks such as Twitter and Reddit have allowed researchers to find new ways to examine the mental health of users. Recent studies have utilized NLP and various machine learning approaches to analyze textual data in posts and provide further insight into depression detection. For example, Shen and Rudzicz [13] collected 22,808 posts from Reddit over three months to study anxiety disorders. Features were generated using N-gram language modeling, vector embeddings, topic analysis, and emotional norms to classify anxiety-related posts [13]. Shen and Rudzicz achieved an accuracy of 98% by combining several NLP techniques. Additionally, Latent Dirichlet Allocation (LDA) topic modeling, a Bayesian generative process, was applied to find the correlations between anxiety and specific topics [13]. The method used by Shen and Rudzicz demonstrates the value of the methodology used in this paper and could be highly effective in identifying depression-related topics that college students encounter.

Machine learning and NLP methodologies have been implemented in various academic fields, such as entrepreneurship. One example is Yu [14], who extracted a corpus of 10,150 posts from Reddit and used machine learning to investigate the areas of struggle in the venture-building process. These struggle-related events can be correlated with depression and other mental illnesses in entrepreneurs. After implementing NLP techniques, the study identified four areas of entrepreneurial struggle: product concept and business model; resources; market entry strategy and timing; and customer care, service, and communication [14]. This study applies a similar machine learning methodology to explore topics related to depression among college students.

Recently, studies have begun to use machine learning to detect depression among college students. Gil et al. [15] examined family data to predict depression in 513 Korean students with three machine learning models: Logistic Regression, Random Forest, and Support Vector Machine. In the study, the Random Forest model had the best performance for depression prediction and was able to identify five factors of depression [15]. Neuroticism, self-perceived mental health, fearful attachment, family cohesion, and depression of parents were correlated with the risk of depression in college students [15]. Gil et al. are one of several studies that show the significance of investigating depression using a machine learning approach. However, the limited research on depression among college students involves restricted datasets. Social media has become a popular resource for detecting depression and other mental illnesses but has not been used in research involving college students. This paper expands on the existing research using computational linguistics and artificial intelligence. It has two objectives: exploring the factors that cause depression among college students and identifying the optimal machine learning classifier for detecting depression in Reddit posts.

3. METHODOLOGY

There is an increasing number of methodologies that can be applied to identify depression. This study uses machine learning classifiers and natural language processing (NLP) techniques for the purpose of depression detection. Figure 1 shows the steps for identifying depression-related factors and the optimal classifier for depression detection.

3.1 Data Collection

Reddit is among the most prominent social networks and is an online discussion forum for thousands of communities known as "subreddits." The platform consists of over 50 million daily active users and more than 100,000 communities [16]. These communities are dedicated to specific topics. This structure enables easy access when people look for posts. Additionally, Reddit allows users to post relatively large bodies of text, and unlike other social media sites, it offers anonymity. The option of anonymity allows honest discussions of stigmatic topics. This study utilizes the anonymity feature to understand depression-related events and problematic issues that college students encounter.

Utilizing the Reddit API, 7,680 semi-anonymous posts were extracted from 13 college-related online forums. Together, the groups consisted of more than 2,660,000 users. The 13 college-related subreddits are shown below:

r/college, r/collegemajors, r/collegerant, r/collegeadvice,
r/collegehub, r/collegeLPT, r/collegestudents, r/collegehelp,
r/highereducation, r/studentaffairs, r/schooladvice, r/school,
r/askacademia

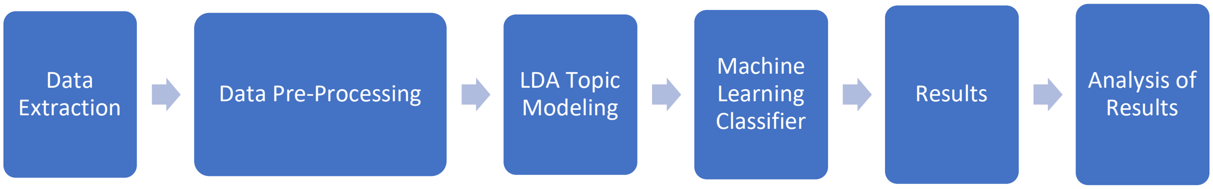

Figure 1: Steps for identifying depression-related events and the most optimal machine learning classifier

A word list consisting of depression-related words was used to filter through the original corpus. It is important to note that the algorithm used in this study included all the verb tenses when appropriate. Below are the base forms of the depression-related words used in the word list:

depression, miserable, sad, unhappy, discourage, melancholy, deject, agonize, gloom, despair, hopelessness, pain, suffer, abject, dispirit

The importance of user anonymity for the present study was confirmed by many of the posts. For example, one college student said: "I struggle with bills, food, and everything. I am lucky enough not to be renting an apartment and never have money for both food and bills. I feel miserable because having no money doesn't help my atmosphere when working on my academic endeavors." Similar sentiments are widely shared across subreddits and viewed by many users.

After applying the word list to the original corpus, 552 posts contained depression-related terms. Before proceeding to topic modeling and prediction, the corpus was pre-processed. Specifically, Python's Natural Language Toolkit [17] library was employed on the dataset. First, URLs, punctuation, and stop words were removed because the results may be erratic and noisy if these are ignored. Furthermore, they do not contribute to the understanding of significant issues under study. Next, lemmatization was conducted on every post to eliminate inflectional endings and retrieve the base form of words. Lemmatization and stemming are both techniques to reduce words to the base form; however, they differ on the level of linguistic analysis. Lemmatization was used instead of stemming because the method uses morphological analysis and ensures that the result is a valid word [18].

3.2 Topic Modeling

After data pre-processing, the posts were converted to numerical data for topic modeling and depression detection. Topic modeling is an unsupervised machine learning technique that clusters words to determine themes across documents and is used in this study to identify the depression-related areas commonly discussed in Reddit posts.

Latent Dirichlet Allocation (LDA) is a generative statistical modeling technique to uncover latent topics within a collection of documents [19]. As LDA is a new approach for identifying depression among college students, its methodology is described here. The generative process of the LDA model is as follows: for each latent topic k in the document collection D, there is a discrete probability distribution $\phi^{(k)}$ over a fixed vocabulary distribution representing the kth topic; for each document $d \in D$, there is a document-specific distribution θ_d over the topics, and for each word $w_i \in d$, z_i is the topic index for w_i; α and β are the hyperparameters for Dirichlet distributions [20]. The process described above results in the joint distribution below:

$$p(W,Z,\theta,\varphi;\alpha,\beta) = \prod_{j=1}^{M} p(\theta_j;\alpha) \prod_{i=1}^{K} p(\varphi_i;\beta) \prod_{t=1}^{N} p(Z_{j,t}|\theta_j)\, p(W_{j,t}|\varphi_{Z_{j,t}})$$

In the above equation, $p(W,Z,\theta,\varphi;\alpha,\beta)$, the left side of the equation, is the probability of generating a particular document. Of the four probability terms on the right side of the equation, the first two represent Dirichlet distributions, and the last two are multinomial distributions [21]. The multinomial distribution is a generalization of the binomial distribution for K outcomes and is shown below for n trials:

$$p(x|\theta) = \frac{n!}{\prod_{i=1}^{d} x_i!} \prod_{i=1}^{d} \theta_i^{x_i}$$

where x_i indicates word i of the vocabulary is observed (0 or 1) and θ_i is the probability that word i is seen $p(w_i)$ [22]. The Dirichlet distribution is a multivariate continuous probability distribution with parameters α, a k-vector, and is conjugate to the multinomial distribution [19][22]. The equation for the Dirichlet distribution is as follows, where Γ() is the gamma function:

$$p(x|\alpha) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}$$

The LDA model was applied to the present dataset to classify depression-related posts into specific topics. In order to use the LDA model approach, the number of topics must be specified. Therefore, the topic coherence measures UCI and UMass are used to assess the quality of the learned topics [23][24][25]. These scores help quantify the interpretability of the topics generated by the model. Both measures compute the coherence score:

$$Coherence = \sum_{i<j} (w_i, w_j)$$

where words w_i and w_j ($w_i \neq w_j$) describe the topic and are scored. The highest coherence measure would return the number of topics to input into the LDA model. The UMass measure is represented in the equation below, where $D(w_i, w_j)$ is the number of documents in which the words w_i and w_j appear together, and $D(w_i)$ is the number of times w_i appeared alone [20]. The value one is added to the numerator to avoid taking the logarithm of 0.

$$score_{UMass}(w_i, w_j) = log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

Similarly, the UCI score, another coherence measure, was calculated. $P(w_i, w_j)$ represents the probability of seeing words w_i and w_j together in a random document, and $P(w)$ is the probability of seeing the word w [24].

$$score_{UCI}(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

After finding the number of topics with the highest coherence scores, an LDA model was trained on the dataset to generate latent unlabelled topics characterized by a selected distribution of the top 5 individual words and N-grams, which form sequences of words.

3.3 Depression Detection

Four classification techniques were employed to estimate the presence of depression among Reddit posts. More specifically, the models tested were the Logistic Regression, Support Vector Machine, Random Forest, and Adaptive Boosting classifiers. Logistic Regression (LR) is a classification approach to predicting a binary outcome based on one or more features. The model is a popular methodology because of its simplicity in classification problems [26]. The Support Vector Machine (SVM) model represents datasets as points on a high dimensional space for classification purposes. The categories of the points are divided by the optimal hyperplane found by the algorithm. Support Vector Machine is effective in high-dimensional spaces and can handle different types of data [27]. Random Forest (RF) is a collection of classification and regression trees that are binary splits of models on variables to predict outcomes [28]. In other words, the algorithm combines the predictions of multiple decision trees to achieve an accurate result. Adaptive Boosting (AdaBoost) classifier combines many weak classifiers to make one robust classifier.

The performance of each machine learning algorithm was evaluated using accuracy, precision, recall, and F1 score. These metrics rely on a confusion matrix, which includes the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) measurements. Accuracy measures the overall correctness of the model's predictions. Precision quantifies the number of positive instances that are correct. Recall calculates the proportion of positive cases that were accurately identified [29]. The F1 score is the harmonic mean of precision and recall. A high F1 score will be obtained if the precision and recall are high. Conversely, a model will have a low F1 score if the precision and recall are low [30]. The accuracy, precision, recall, and F1 score equations are shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2\frac{precision \times recall}{precision + recall}$$

4. RESULTS AND DISCUSSION

4.1 Topic Modeling Results and Discussion

The coherence measures allow for straightforward and quick tests to see the number of topics that should be assigned to the LDA model. The UMass and UCI coherence measures were calculated for topics ranging from two to nine. For the UMass evaluation, the highest coherence value was -1.71 when the number of topics was three. In Figure 2, the results for the UMass coherence score are shown accordingly.

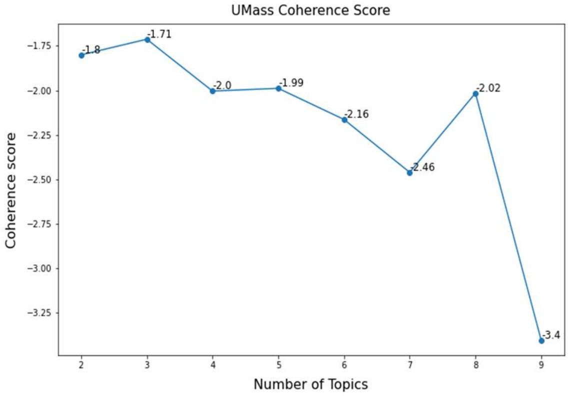

Figure 2: UMass Coherence Score

As the coherence score of -1.71 is similar to the scores of two, four, and five topics, the UCI metric helped confirm the UMass measure. The UCI measurement had the highest coherence score at -1.09 when the number of topics was three. It is also important to note that the UCI coherence score when the number of topics was three was considerably higher than any other topic, with the nearest score at -1.63. The coherence scores for the UCI metric are presented in Figure 3.

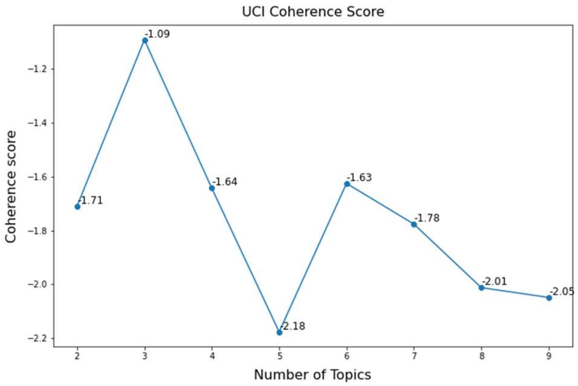

Figure 3: UCI Coherence Score

Taking into account that the UMass and UCI coherence scores are the highest when the number of topics is 3, the LDA model was trained to classify the depression-related posts into three categories. Figure 4 shows the T-SNE [31] clusters of three topics to represent the topic arrangements visually.

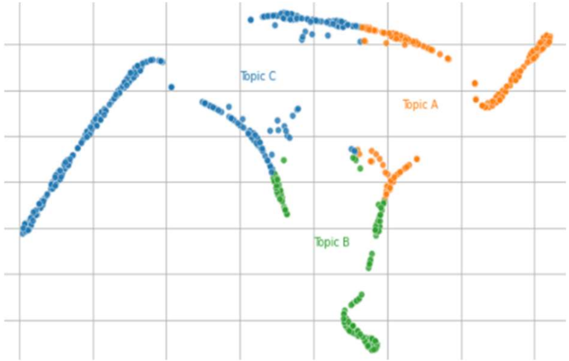

Figure 4: T-SNE topic representation

Of the 552 posts containing depression-related words, 165, 110, and 277 were assigned to Topic A, B, and C, respectively. Figure 5 conveys the distribution of depression-related posts graphically.

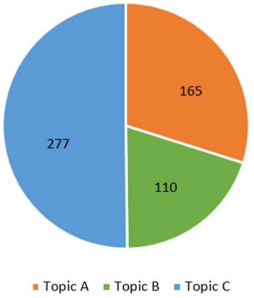

Figure 5: Distribution of topics

Table 1 presents the results of the LDA model for the three topics. The topic words and N-grams were related to institutions and programs for Topic A. Depression-related posts assigned to Topic A indicate that students are unhappy or dejected about their current college program or major. The majority of the top words and N-grams of Topic B expressed frustrations with academic projects and assignments. In other words, posts assigned to Topic B reflected discouragements about assignments or feedback left by a professor. Topic C contains words and N-grams about the college environment in general. Under Topic C, users may feel uncomfortable in the college setting.

Table 1: Top words/N-grams for each topic

	Top 5 Topic Words/N-grams
Topic A	community college, high school, career, university, science
Topic B	position, academic, professor, need help, paper
Topic C	high school, don't know, mental health, grade, first semester

Based on the findings of this study, universities and colleges should focus on three primary depression-related areas: institutions and programs; academic projects and assignments; and the college environment. Providing resources to college students so they can adjust to the college environment and overcome depression-inducing events is critical, as over half of the depression-related posts fell under Topic C. Examples of depression-related posts from each topic are provided below to help illustrate the themes of the topics. Topic A, related to institution and program, is evident in the post:

My major is honestly so discouraging… I feel like it's too late to switch again, and honestly, I don't want to. Is it worth it to get masters in something different? Like Child psychology or something of that matter?

The college student in the post above is discouraged by their current major and reached out to Reddit to discuss the issue. Many posts assigned to Topic A had similar sentiments about the students' majors or programs. Posts in Topic B expressed frustrations with academic projects and assignments. The user who posted the message below is discouraged by the comments left on a recent assignment by the instructor:

So, I'm in this major-specific course, and I feel like it's been going well so far. Well, recently, I've been getting feedback on my assignments, and I didn't expect the negative feedback I received. A specific comment stood out to me, saying, "I should have known better" than to choose a source that wasn't what the assignment was looking for. Everyone else in the class is at a higher grade than I am since it's considered an upper-level course, but my professor has stated that they consider me a "senior-level student." I'm glad they take pride in me and have encouraged me throughout my time in school, but now I'm just feeling particularly discouraged and insecure. I feel like all these higher-level students are doing much better than me and that I could be doing much better even though this might be my best."

Complaining about the college environment was reflected in many posts assigned to Topic C. Very often, the college environment exhausted students, leading them to question if college was right for them. This can be seen in the post below:

Am I supposed to be enjoying college? I'm miserable, but everyone around me is having a better time. I hear people having fun and doing work together in the dorms next to me. I'm doing so bad, though. I'm passing my classes, but I feel like the work I do is draining me so much. I have no time for anything else.

4.2 Depression Detection Results and Discussion

The performance metrics of the classification models are shown in Table 2. Of the algorithms, the AdaBoost classifier had the highest accuracy at 98.984%, followed by Support Vector Machine (94.674%), Random Forest (93.529%), and Logistic Regression (93.307%). However, the Random Forrest classifier could predict depression with a 1.000 F1 score, compared to the 0.957 F1 score achieved by the AdaBoost classifier. It is also important to note that Support Vector Machine obtained an accuracy of 94.674% with a near-perfect 0.990 F1 score. The Logistic Regression classifier achieved an accuracy of 93.307% and an F1 score of 0.755. The confusion matrices for Logistic Regression, Support Vector Machine, Random Forest, and AdaBoost are displayed in Figures 6, 7, 8, and 9, respectively.

Although the AdaBoost classifier attained the highest accuracy, the Random Forest classifier had a greater F1 score. As the F1 score metric considers the model's precision and recall, it provides a more comprehensive evaluation [32]. In the future, the Random Forest classifier should be used for datasets with more variation, while the AdaBoost classifier may be chosen for datasets with smaller amounts of variation.

With a statistically accurate approach to detecting depression among college students, this study contributes to broader academic research through a novel machine learning methodology. The machine learning approach presented here shows promising accuracy and precision and has potential for clinical practice. As the number of people with depression increases daily, an automated system is required to detect early signals of depression and treat patients accordingly. Interviews are often conducted, and traditional diagnosis is costly [33]. Adequate depression diagnosis and treatment may require several weeks, and machine learning methods can significantly reduce the duration that the patient suffers [34]. In addition, these techniques have wide-ranging implications and can be applied to various psychiatric diseases, such as schizophrenia and obsessive-compulsive disorder.

In past depression prediction research, small sample sizes and various learning methods have been used [34]. This study incorporates several machine learning algorithms and a more extensive database than previous research. Furthermore, the results of this study indicate that online forums can be used for interventions, as many posts suggest that college students find it assuring to have a supportive community to share mental health issues without the need for protecting privacy and image. This can lead to safe discussions about depression and other mental health issues not offered in anonymous and stigmatized environments. Therefore, efforts at integrating machine learning techniques and online forum databases contribute to the efficiency of detecting depression among college students and can help prevent future mental health issues.

Table 2: Performance results of the classification models

	LR	SVM	RF	AdaBoost
Accuracy	93.307%	94.674%	93.529%	98.984%
F1 Score	0.755	0.990	1.000	0.957
Precision	0.956	1.000	1.000	1.000
Recall	0.844	0.995	1.000	0.917

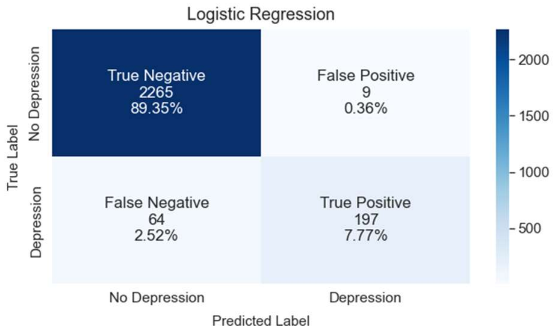

Figure 6: Logistic Regression confusion matrix

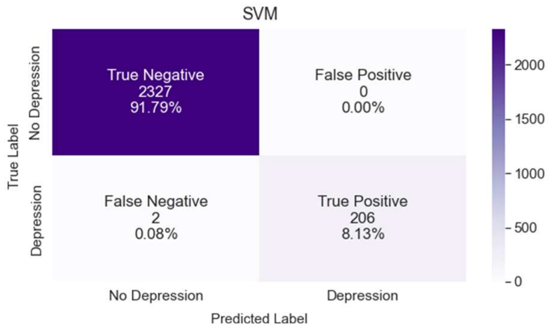

Figure 7: Support Vector Machine confusion matrix

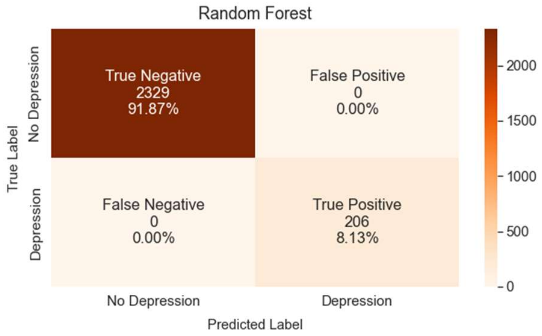

Figure 8: Random Forest confusion matrix

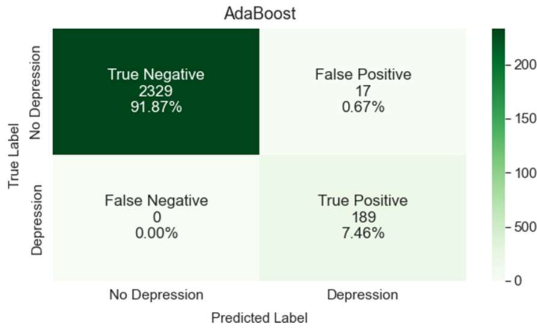

Figure 9: AdaBoost confusion matrix

5. LIMITATIONS AND FUTURE WORK

The findings in this study afford many directions and implications for future academic research. As the machine learning methodology continues to be implemented in the detection of psychiatric illnesses, progressing to clinical trials to evaluate whether the machine learning methodology will benefit patients is a promising step toward automated diagnosis and treatment. Although the posts in the corpus contained large amounts of textual data, audio, image, and video features were not considered when identifying depression-related areas or detecting depression. Further research is required to implement a combination of textual, image, audio, and video components for potentially optimal depression detection.

While the usefulness of machine learning and online forum databases are acknowledged in this study, it is important to recognize that the methodology cannot yet replace the insights and analysis of clinical professionals. Moreover, the corpus does not contain detailed information about the users, and Reddit participants are not representative of the general population. According to a recent Reddit user statistics report, 74% of users are male compared to 25.8% female, and over 60% are white [35][36]. Additionally, 44% of users identify as liberal, while only 19% are conservatives (38% are moderate) [36]. Thus, the experiences of college students with different characteristics may not be fully captured. As this research aimed to minimize identity disclosure, all participants were anonymized, and only the body text of posts was extracted. This excludes personal data such as the author's username, which protects the user's identity.

In addition, the circumstances and mood in which college students post could influence the contents of the posts. For example, some students could have caught a cold or had a poor night's rest when they posted. Such information would improve the accuracy of the machine learning algorithm and be relevant for the appropriate prevention and intervention strategies. Future research can make an effort to identify these cases and their effect on depression.

Although depression-related areas among college students were identified, future studies should make efforts to research the magnitude of the impact of various depression factors on students. While Reddit allows users to post stigmatic topics, it does not let researchers examine the previous experiences of college students. These previous backgrounds may reveal how college students handle depression-related events and move on from them. In the future, variables such as the severity of depression-related experiences should be taken into account when detecting depression among college students.

6. CONCLUSION

In this study, three primary areas linked to depression among college students were identified: institutions and programs; academic projects and assignments; and the college environment. Of the 7,680 college-related Reddit posts, 552 contained keywords of depression. It is essential to use the findings of this research to create educational materials and programs to help college students better prepare for the inevitable trials and tribulations during their college years. Organizations and educational institutions should especially focus on the depression-related areas identified in this study to help college students cope with depression-related issues.

For detecting depression among college students, the Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Adaptive Boosting (AdaBoost) classifiers were utilized. The AdaBoost classifier successfully predicted depression in Reddit posts with the highest accuracy at 98.984%, and the Random Forrest classifier detected depression at a 1.000 F1 score. Furthermore, the Support Vector Machine algorithm had a 94.674% and a 0.990 F1 score. This study shows that machine learning is a promising methodology for detecting depression in Reddit posts and can help determine the areas that induce depression in college students.

7. REFERENCES

[1] American Psychiatric Association. (2020, October). *What is depression?* Psychiatry.org – What is Depression? Retrieved March 20, 2023, from https://www.psychiatry.org/patients-families/depression/what-is-depression

[2] Mayo Clinic Health System. (2023, May 31). *College students and Depression*. Mayo Clinic Health System. https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/college-students-and-depression

[3] National Institute of Mental Health. (2020). *Major depression*. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/major-depression#:~:text=In%202020%2C%20an%20estimated%2066.0,treatment%20in%20the%20past%20year

[4] Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders*, *173*, 90–96. https://doi.org/10.1016/j.jad.2014.10.054

[5] Thurber, C. A., & Walton, E. A. (2012). Homesickness and adjustment in university students. *Journal of American College Health*, *60*(5), 415–419. https://doi.org/10.1080/07448481.2012.673520

[6] Sun, J., Hagedorn, L. S., & Zhang, Y. (Leaf). (2016). Homesickness at college: Its impact on Academic Performance and Retention. *Journal of College Student Development*, *57*(8), 943–957. https://doi.org/10.1353/csd.2016.0092

[7] Barbayannis, G., Bandari, M., Zheng, X., Baquerizo, H., Pecor, K. W., & Ming, X. (2022). Academic stress and mental well-being in college students: Correlations, affected groups, and covid-19. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.886344

[8] Liu, X. Q., Guo, Y. X., Zhang, W. J., & Gao, W. J. (2022). Influencing factors, prediction and prevention of depression in college students: A literature review. *World journal of psychiatry*, *12*(7), 860–873. https://doi.org/10.5498/wjp.v12.i7.860

[9] Goswami, S., Sachdeva, S., & Sachdeva, R. (2012). Body image satisfaction among female college students. *Industrial psychiatry journal*, *21*(2), 168–172. https://doi.org/10.4103/0972-6748.119653

[10] Orzech, K. M., Salafsky, D. B., & Hamilton, L. A. (2011). The state of sleep among college students at a large public university. *Journal of American college health: J of ACH*, *59*(7), 612–619. https://doi.org/10.1080/07448481.2010.520051

[11] Doom, J. R., & Haeffel, G. J. (2013). Teasing apart the effects of cognition, stress, and depression on health. *American Journal of Health Behavior*, *37*(5), 610–619. https://doi.org/10.5993/ajhb.37.5.4

[12] Ebert, D. D., Buntrock, C., Mortier, P., Auerbach, R., Weisel, K. K., Kessler, R. C., Cuijpers, P., Green, J. G., Kiekens, G., Nock, M. K., Demyttenaere, K., & Bruffaerts, R. (2018). Prediction of major depressive disorder onset in college students. *Depression and Anxiety*, *36*(4), 294–304. https://doi.org/10.1002/da.22867

[13] Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through Reddit. Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality. https://doi.org/10.18653/v1/w17-3107

[14] Yu, P. (in press). Entrepreneurial Struggle: A Natural Language Processing Approach. *International Journal of High School Research*.

[15] Gil, M., Kim, S.-S., & Min, E. J. (2022). Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1023010

[16] Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and Ethics. *Social Media + Society*, *7*(2), 205630512110190. https://doi.org/10.1177/20563051211019004

[17] Loper, E., & Bird, S. (2002). NLTK. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics -. https://doi.org/10.3115/1118108.1118117

[18] Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, *2*(3), 262–267. https://doi.org/10.7763/lnse.2014.v2.134

[19] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[20] Darling, W. M. (2011, December). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 642-647).

[21] GeeksforGeeks. (2021, June 6). *Latent dirichlet allocation*. GeeksforGeeks. https://www.geeksforgeeks.org/latent-dirichlet-allocation/

[22] Clark, S. (2013). Topic modelling and latent dirichlet allocation. *Online, Lent*.

[23] Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014, March 25). *Evaluating topic coherence measures*. arXiv.org. https://arxiv.org/abs/1403.6397

[24] Zvornicanin, W. by: E. (2023, May 31). *When coherence score is good or bad in topic modeling?*. Baeldung on Computer Science. https://www.baeldung.com/cs/topic-modeling-coherence-score

[25] Pleplé, Q. (2013). *Topic Coherence To Evaluate Topic*

Models. Topic coherence to evaluate topic models. http://qpleple.com/topic-coherence-to-evaluate-topic-models/

[26] Edgar, T. W., & Manz, D. O. (2017). Science and cyber security. *Research Methods for Cyber Security*, 33–62. https://doi.org/10.1016/b978-0-12-805349-2.00002-9

[27] Noble, W. S. (2006). *What is a support vector machine?*. Nature News. https://www.nature.com/articles/nbt1206-1565

[28] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

[29] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit Social Media Forum. *IEEE Access*, *7*, 44883–44893. https://doi.org/10.1109/access.2019.2909180

[30] Korstanje, J. (2021, August 31). *The F1 score*. Medium. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6#:~:text=The%20F1%20score%20is%20defined,when%20computing%20an%20average%20rate.

[31] van der Maaten , L., & Hinton, G. (2008). *Visualizing data using T-SNE*. Journal of Machine Learning Research. https://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[32] Huilgol, P. (2019, August 24). *Accuracy vs. F1-score*. Medium. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

[33] Vandana, Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using Deep Learning. *Measurement: Sensors*, *25*, 100587. https://doi.org/10.1016/j.measen.2022.100587

[34] Patel, M. J., Khalaf, A., & Aizenstein, H. J. (2016). Studying depression using imaging and Machine Learning Methods. *NeuroImage: Clinical*, *10*, 115–123. https://doi.org/10.1016/j.nicl.2015.11.003

[35] Gitnux, A. (2023, July 12). *Reddit user statistics and Trends in 2023 • gitnux*. GITNUX. https://blog.gitnux.com/reddit-user-statistics/#:~:text=engage%20in%20conversations.-,With%20over%20430%20million%20monthly%20active%20users%2C%2074%25%20of%20which,ranging%20from%20politics%20to%20entertainment.

[36] Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016). Seven-in-ten Reddit users get news on the site.