Analyzing Suicidal Ideation in Online Communities using Natural Language Processing Techniques

Sadid Islam, Mubashira Rahman,
Riead Hasan Khan,
Department of Computer Science and Engineering
Brac University, Dhaka, Bangladesh
Email: sadid.islam@g.bracu.ac.bd, mubashira.rahman@g.bracu.ac.bd , md.riead.hasan@g.bracu.ac.bd

***Abstract*—This study uses several machine learning models to analyze posts from the r/SuicideWatch subreddit in order to identify patterns associated with suicidal ideation. Our contribution includes a comparative analysis of machine learning models for the task of suicidal ideation classification based on textual data. For a dataset as large as ours, instead of focusing on transformer-based models like BERT or deep learning models, we focused on traditional machine learning models and achieved acceptable 93.22% accuracy with the SVM classifier. Additionally, this paper highlights the potential of social media platforms as a valuable source of data for studying suicidal ideation and the effectiveness of machine learning techniques for analyzing large volumes of data in this area of research.**

***Index Terms*—Suicide ideation, Natural Language Processing, Support Vector Machine Classifier, Multinomial Naive Bayes Classifier, Random Forest, and Multilayer Perceptron Classifier, social media.**

I. Introduction

Suicide has become a global public health crisis, affecting millions worldwide. Suicide rates are on an upward trajectory in numerous regions of the world despite growing awareness and attempts at tackling this global problem, causing it to be a significant and continuous concern for medical professionals, legislators, and mental health activists. Many more individuals try suicide each year, in addition to the 703,000 suicides that occur. Suicide is an awful event that affects the survivors, their families, communities, and whole countries, deeply. Suicide may happen to anybody at any age, and in 2019, it was the fourth leading cause of death globally for those between the ages of 15 and 29 [1]. Finding and comprehending the components that lead to suicidal thoughts is an essential field of research since it is a severe issue that needs attention and help.

Reddit and other social media sites have emerged as significant data sources for scholars looking at this topic. The relationships between general SNS usage and higher levels of teenage ill-being found in five meta-analyses ranged from extremely low to substantial [2]. By examining the language and themes covered in posts, researchers might learn more about the causes of suicidal thoughts and spot those who might be at risk for engaging in suicidal conduct. Utilizing machine learning approaches has greatly improved our capacity to examine and comprehend information from social media that relate to suicidal thoughts. Even when people are not declaring their intention to kill themselves, these strategies can help researchers find linguistic patterns that are suggestive of suicidal thoughts or acts. The research of suicidal thoughts on social media sites is becoming more common, and machine learning techniques like sentiment analysis and natural language processing (NLP) are being employed more often. With the use of these tools, enormous volumes of data may be quickly and accurately analyzed, yielding insightful information on the emotional states and linguistic patterns connected to suicidal thoughts.

In this study, we aim to use several machine learning models, such as the Multinomial Naive Bayes Classifier, Random Forest Classifier, Support Vector Classifier and finally Multilayer Perceptron Classifier to evaluate posts from the r/SuicideWatch subreddit, a community where people may share their suicide ideation difficulties and get help and assistance. Many deep learning models and transformer architectures have been proven to be very accurate in this field of classification, but they are quite resource-intensive. So, our goal is to find the effectiveness of Multinomial Naive Bayes Classifier, Random Forest Classifier, Support Vector Classifier and finally Multilayer Perceptron Classifier for the task of suicidal ideation detection as an alternative to resource-intensive processes. .

II. Related works

The combination of the data from social media with machine learning and NLP offers a singular chance to spot and help people who are at risk of suicide and has the potential to have a big influence on efforts to prevent suicide. Researchers can create more potent suicide detection and response efforts by spotting these tendencies in social media data.

In a study Agarwal and Dhingra present a novel approach based on deep learning to detect suicidal intent in literature featuring Hindi-English code mixing. Their study introduces the Hinglish Suicide Ideation Corpus (HSIC), a unique collection of social media posts that combine Hindi and English languages and have been annotated for suicidal ideation. The proposed technique utilizes a deep neural network architecture, incorporating contextual information, semantic and syntactic elements, and language embeddings, to determine the presence of suicide intent in each post. Results indicate that code-mixed posts achieve an overall F1 score of 0.55, while English and

Hindi posts achieve F1 scores of 0.60 and 0.50, respectively [3].

Another paper suggests a unique method for identifying suicidal thoughts in Chinese microblogs using psychological lexicons. After manually annotating a dataset of Chinese microblogs from Sina Weibo, they used a set of psychological lexicons to extract semantic features from the text to find microblogs showing suicidal ideation. They then compared their method to other algorithms to gauge its effectiveness. This strategy works well at spotting people in danger of acting suicidally and giving them the right treatments. The psychological lexicon-based strategy fared better than the other strategies, as evidenced by its F1 score of 0.72 [4].

According to Yeskuatov and his team, because of the unstructured and cluttered nature of the material, detecting self-harm posts on digital platforms is a difficult process [5]. They have suggested machine learning algorithms and NLP methods to deal with this problem. This work developed a classifier using mood, emotion, and language variables using a dataset of Reddit postings on suicide. The F1-score of 0.83 obtained by the classifier indicates great accuracy. This classifier's F1-score of 0.86 demonstrates the strategy's success. The strategy outperformed conventional machine learning methods with an F1-score of 0.87. In all, it has been discovered that a mixture of lexical and semantic data, together with deep learning-based methodologies, is efficient in identifying suicidal intent on Reddit. Another research by Ji.S et al. offers an overview of machine learning techniques and programs for identifying suicidal thoughts on social media sites [6]. The authors looked at papers that suggested several strategies based on multimodal, linguistic, and semantic aspects. They highlighted several possible applications for suicide prevention and intervention. High accuracy and sensitivity were among the positive traits, whereas low specificity and generalizability were the negative traits. The article suggests a unique method for identifying suicidal thoughts on Twitter by using several feature analyses [7]. In order to extract information from a collection of tweets involving suicidal thoughts, the authors employed linguistic, semantic, sentiment, network, and temporal aspects. The suggested method successfully identified suicidal thoughts on Twitter with excellent accuracy, sensitivity, and specificity. After comparing it to those approaches, the authors stated that their approach outperformed existing machine learning techniques.

The study by Patel et al. analyzes internet health discussion boards to find mental and physical problems connected to COVID-19. According to the report, the two most prevalent mental ailments are anxiety and depression, while the three most prevalent physical disorders are exhaustion, respiratory issues, and loss of taste and smell. The authors emphasize the potential of NLP methods for examining digital healthcare communities and spotting linguistic patterns and trends among forum users. The study may not reflect the opinions and experiences of the larger public because it relies solely on the information provided by individuals from online health forums [8]. Kabir et al. detected depression severity from Bengali social media texts using natural language processing techniques [9]. They used text-based data in Bengali from blogs and open postings to create a procedure for creating annotated corpora and fetched data. They utilized DSM-5 to properly diagnose and Selenium to scrape data from social media. They employed four separate classified labels to distinguish between the texts. They employed models such as the random forest, support vector machine, logistic regression, k-nearest neighbor, and Naive Bayes for pre-processing and data modeling. In comparison to previous models, the authors claim that the recurrent neural network model successfully predicted the severity [10]. Priya et al. discuss the purpose of the study, data were gathered by utilizing the DASS-21 questionnaire and text analysis from social media. They employed various machine learning algorithms and evaluated which one provided the highest level of accuracy. For instance, when they used the CNN and SVM algorithms to forecast stress and anxiety, the CNN algorithm performed with an accuracy of around 79% whereas the SVM algorithm performed with only 58% [11]. They did this by comparing each algorithm's output for each prediction and selecting the algorithm with the best accuracy for predicting the outcome.

Lastly, another study used assessments of eight fundamental emotions as characteristics from Twitter tweets throughout time, including a temporal analysis of these features, to create a technique. They measured emotions using the EMOTIVE technology and Ekman's core emotion model. For differentiation, the data set was divided into sets of temporal and nontemporal features. Statistical and mathematical techniques were applied, including mean, standard deviation, entropy, mean momentum, and mean differencing. They contrasted the outcomes of temporal and nontemporal data sets generated by the suggested systems [12].

III. Dataset

The dataset compiles postings from "Reddit" by "SuicideWatch" and "depression" subreddits. The Pushshift API is used to gather the posts. All posts made to "SuicideWatch" between Dec. 16, 2008 (when it was created) and Jan. 2, 2021 were gathered, whereas postings made to "depression" between Jan. 1, 2009, and Jan. 2, 2021, were gathered [13]. This dataset has 232074 unique values divided and labeled into 2 categories, suicide or non-suicide. The texts are basically people sharing their feeling publicly on social media. The mention of suicide was evident in the texts. We used 20 thousand data from this dataset in this research

IV. Data Prepocessing

The dataset was divided into testing and training sets with a fixed random state of 42, preserving an 80:20 ratio for consistency.The data was tokenized and vectorized using 'sklearn' library. Different deep learning and machine learning models, such as Multi-layered Perceptron Classifier Random Forest, Support Vector Classifier, and Multinomial Naive Bayes were

trained using the training data. To improve each model's performance, appropriate hyperparameters were added.

V. Methodology

A. Model Training and Evaluation

The dataset was divided into testing and training sets with a fixed random state of 42, preserving an 80:20 ratio for consistency. Different deep learning and machine learning models, such as Multi-layered Perceptron Classifier Random Forest, Support Vector Classifier and Multinomial Naive Bayes were trained using the training data. To improve each model's performance, appropriate hyperparameters were added.

The quantity of the dataset made it unfeasible to use an automated method to find the best hyperparameters due to the length of time it would take. An alternative method involved manually adjusting the hyperparameters and assessing how it affected the model's accuracy was chosen. By identifying local accuracy maxima using this technique, the most effective hyperparameters might be chosen.

B. Multinomial Naive Bayes

We have set the parameter 'alpha' as 0.1 Multinomial Naive Bayes classifier and denoted 'fit prior' as true. After the algorithm was tuned to the training dataset, predictions on the test dataset were obtained.

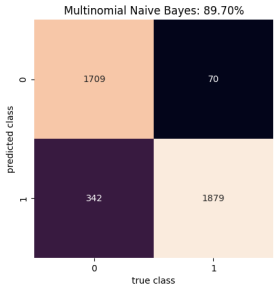

Figure 1: (a) Confusion Matrix of Multinomial Naive Bayes

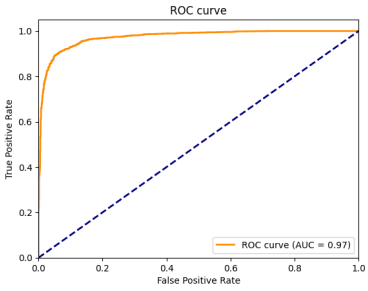

Figure 1: (b) ROC curve of the Multinomial Naive Bayes

As seen in Figure 1, (a) shows the confusion matrix, and (b) for the ROC curve of Multinomial Naive Bayes classifier ROC curve. In the test dataset, 89.7% of the cases had their labels predicted correctly. 90.47% of the anticipated positive events were accurately classified. The model's high level of accuracy in differentiating between positive and negative classifications is demonstrated by the area under the curve (AUC), which is 0.97.

C. Support Vector Classifier

We have used the "linear" kernel parameter to build the SVM classifier $c = 1$ is set as the regularization value The "probability" parameter was also enabled by choosing "True." Reproducibility was ensured by setting the "random state" parameter to 42.

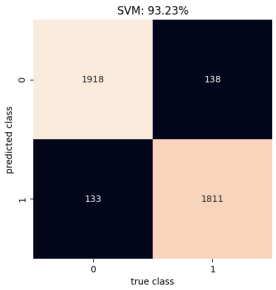

Figure 2: (a) Confusion Matrix of Support Vector Classifier

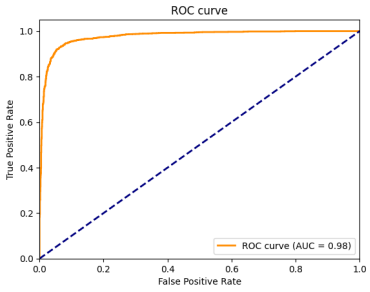

Figure 2: (b) ROC curve of Support Vector Classifier

The results are shown in Figure 2, where (a) is the confusion matrix and (b) is the support vector classifier's ROC curve. The Support Vector Machine classifier's area under the curve is calculated to be 0.98. The classifier's overall accuracy was 93.22%.

D. Random Forest Classifier

For the Random Forest Classifier, we set the number of trees to 500, the maximum depth of each tree to 100, the minimum number of samples needed at a leaf node to 2, and the random seed value to 42. The training data were then fitted to the Random Forest classifier.

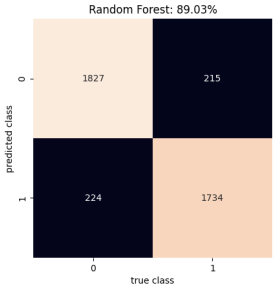

Figure 3:(a) Confusion Matrix of Random Forest Classifier

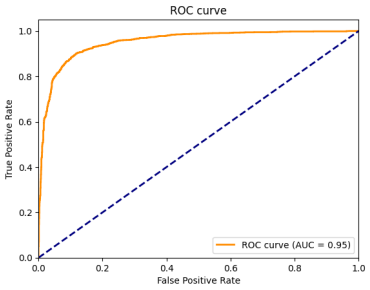

Figure 3:(b) ROC curve of Random Forest Classifier

As seen in Figure 3, (a) represents the ROC curve and (b) represents the Confusion matrix. The area under the curve score for the Random Forest classifier is 0.95. The classifier had an overall accuracy of 0.8902, which means it correctly predicted the label of 89.02% of the cases in the test dataset.

E. Multilayer Perceptron classifier

For the Multilayer Perceptron classifier, we used a hidden layer with 50 neurons and the 'ReLU' activation function. We also used the 'adam' solver. For L2 regularization, the classifier employed an alpha value of 0.0001 and an initial learning rate of 0.001. The random state was set to 42, and the maximum number of iterations was set at 500.

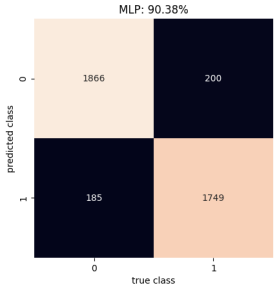

Figure 4: (a) Confusion Matrix of Multilayer Perceptron classifier

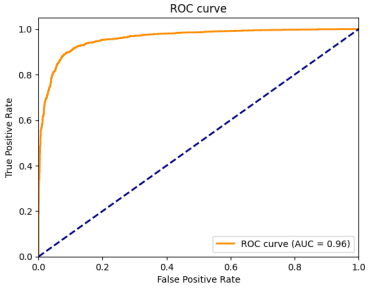

Figure 4: (b) ROC curve of Multilayer Perceptron classifier

As shown in Figure 4, (a) represents the confusion matrix, and (b) represents the ROC curve. The accuracy of the MLP classifier is 0.9037, or 90.37%. The area under the curve score is 0.96.

VI. Discussion

In this research, we have used precision, recall, F1 score for generating the output. After using the algorithms the results were compared. The comparison between the results of algorithms gave much clarity to the work. In the confusion matrix "true class" was plotted on the X axis whereas "predicted class" was plotted on the Y axis. For the ROC curve graph, X axis was "False positive rate" and Y axis was "True positive rate".

TABLE I
PERFORMANCE EVALUATION OF ALGORITHMS

	Precision	**Recall**	**F1 score**
Accuracy	0.93225	0.93225	0.93225
Macro avg	0.932232	0.932174	0.932201
Non-suicide	0.960652	0.935154	0.934015
Suicide	0.931584	0.964084	0.930388
Weighted avg	0.932248	0.93225	0.932248

For all the algorithms the performance evaluation is represented in the table. The best result for the evaluation matrix shows the performance evaluation. Finally plotting the results of all the algorithms can be clearly seen the differences. The slopes were compared against a perfectly calibrated curve. In the X axis mean productive value and in Y axis fraction of positives were plotted.As seen from the calibration curve Support Vector Machine Classifier had the best performance with an accuracy of 93.22% .

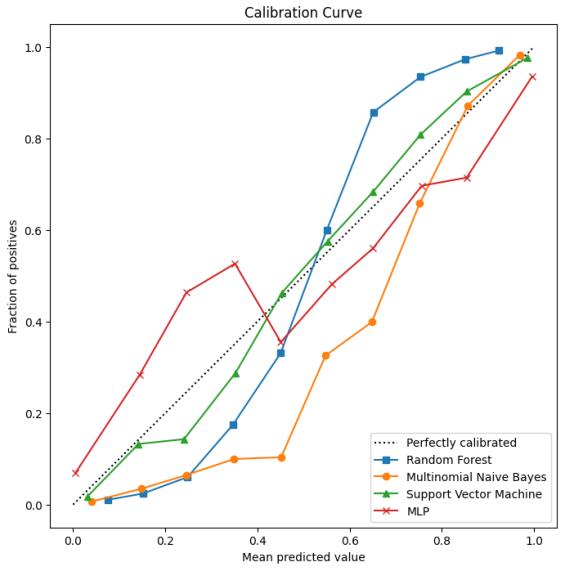

Figure 5: Comparison of all algorithms

VII. Future work and Conclusion

A. Future Work

To improve upon our work, future work should focus on creating a larger dataset. Also, working with a multilingual dataset is more contextual, given that most of the posts are written in mixed languages. This would make it adaptable to different languages. Furthermore, improvements can be made

by evaluating the performance of transformer models and other deep learning models in terms of accuracy, resource consumption, and training time and then comparing them with the models used in our research. So that we can identify the strengths and weaknesses of the models involved in creating a more efficient and accurate model for suicidal ideation detection.

B. Conclusion

In conclusion, social media sites have become part and parcel of our lives, and with people venting their frustrations online, they have become a valuable source of data. With the increasing rate of suicide and social media platforms in mind, we worked to create an accurate model for suicidal ideation detection. Our research involved four different models- Support Vector Machine Classifier, Multinomial Naive Bayes Classifier, Random Forest and Multilayer Perceptron Classifier. Which Support Vector machine had an accuracy of 93.22%. The other models also had good accuracy. Considering that none had an AUC score below 95% on the classification task we can see that these models are an alternative to more resource-hungry and time-consuming models involving transformer architectures and neural networks. Overall, our study provides valuable insights into the strengths and weaknesses of traditional machine learning algorithms for the task of suicidal ideation classification.

REFERENCES

[1] WHO, "Suicide," Who.int, 06 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] Z. Vahedi and L. Zannella, "The association between self-reported depressive symptoms and the use of social networking sites (sns): A meta-analysis," *Current Psychology*, vol. 2, 01 2019.

[3] K. Agarwal and B. Dhingra, "Deep learning based approach for detecting suicidal ideation in Hindi-English code-mixed text: Baseline and corpus," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 100–105. [Online]. Available: https://aclanthology.org/2021.icon-main.14

[4] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting suicidal ideation in chinese microblogs with psychological lexicons," in *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, Dec 2014, pp. 844–849.

[5] E. Yeskuatov, S.-L. Chua, and L. K. Foo, "Leveraging reddit for suicidal ideation detection: A review of machine learning and natural language processing techniques," *International journal of environmental research and public health*, vol. 19, no. 16, p. 10347, 2022.

[6] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, Feb 2021.

[7] M. Chatterjee, P. Samanta, P. Kumar, and D. Sarkar, "Suicide ideation detection using multiple feature analysis from twitter data," in *2022 IEEE Delhi Section Conference (DELCON)*. IEEE, 2022, pp. 1–6.

[8] R. Patel, F. Smeraldi, M. Abdollahyan, J. Irving, and C. Bessant, "Analysis of mental and physical disorders associated with covid-19 in online health forums: a natural language processing study," *BMJ open*, vol. 11, no. 11, p. e056601, 2021.

[9] M. K. Kabir, M. Islam, A. N. B. Kabir, A. Haque, and M. K. Rhaman, "Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques," *JMIR Formative Research*, vol. 6, no. 9, p. e36118, 2022.

[10] L. J. Cohen, S. Wilman-Depena, S. Barzilay, M. Hawes, Z. Yaseen, and I. Galynker, "Correlates of chronic suicidal ideation among community-based minor-attracted persons," *Sexual Abuse*, vol. 32, no. 3, pp. 273–300, 2020.

[11] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020.

[12] X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, "What about mood swings: Identifying depression on twitter with temporal measures of emotions," in *Companion proceedings of the the web conference 2018*, 2018, pp. 1653–1660.

[13] "Suicide and depression detection," www.kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch?fbclid=IwAR2-YeHlYDZc0w2nLZ1k_GYyXOPrBrmSNWSXebNr-SbFd8CgFnK6q5dFNr8