2024 27th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

SuicideGuard: An NLP-Based Chrome Extension for Detecting Suicidal Thoughts in Bengali

1st Nahida Fatme
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
nahida.nine@gmail.com

2nd Natasha Tanzila Monalisa
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh
natasha.tanzila786@gmail.com

3rd Rashedul Jisan
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
rashedjisan@gmail.com

4th Md. Tahsin Rahman
Department of Computer Science and Engineering
North South University
Dhaka, Bangladesh
jishan151297@gmail.com

5th Sanjida Akter
Department of Computer Science and Engineering
Northern University of Bangladesh
Dhaka, Bangladesh
sanjida.nub@gmail.com

6th Shinthi Tasnim Himi
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
shimi@bubt.edu.bd

2024 27th International Conference on Computer and Information Technology (ICCIT) | 979-8-3315-1909-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCIT64611.2024.11022363

***Abstract*—The rate of suicide due to depression among youngsters is ever on the rise. As they are the highest users of social media like Facebook and Twitter, they tend to share their feelings on those platforms. Even during mental breakdowns or while having thoughts about self-harm, they often post statuses that sometimes reflect their inner emotions but are mostly neglected by friends and families. This paper develops a chrome extension 'SuicideGuard' which is a groundbreaking solution to identify suicidal thoughts. It is a tool to help Bengali-speaking individuals by mining out depressions from their posts, as it has been trained on 2,590 Bangla data. This immensely useful system has been trained with the BERT model with 92% accuracy after analysing models like BiLSTM, and XLM-RoBERTa which can accurately predict suicidal thoughts in real-time and will potentially save lives. It will also be beneficial for a psychiatrist as it helps to understand the severity of their patient's anxiety, and even aid the individuals to support their friends and family going through mental instability.**

***Index Terms*—Suicide, LLM, Extension, Mental Health, Social Media**

I. Introduction

According to the World Health Organization (WHO), every year more than 700,000 people die by suicide [1]. This increasing suicidal rate is mostly due to declining mental health [2]. It has been seen that 50% of mental health issues are developed by the age of 14 and 75% by the age of 24 [3]. This proves why suicide is the fourth leading cause of death among 15 to 29-year-olds [1]. Social media like Facebook ($3.06 billion) and Instagram ($2.35 billion) have a huge user base. People use these platforms to post or comment about their day-to-day life activities, achievements, failures, and ultimately their mental state [4]. As of February 2024 reports, the biggest group of users of this media are aged between 18 to 24 [5], which is the same age bracket of peak suicidal rate. Consequently, suicide after posting on social media is reported in different parts of the world. The study shows that all the deceased were under 35 and they shared a series of posts mentioning specific suicidal ideas several times before committing suicide. It seemed that their friends treated them brutally and made fun in the posts' comment section rather than helping them [6]. If they were taken seriously, maybe those lives could have been saved. As those suicidal posts motivate other people to take such steps, amid the onslaught of lawsuits, Instagram and Facebook declared they would start hiding posts about suicide and self-harm [7]. But sometimes there are no direct mentions of suicidal terms, their simple words speak about their mental health and urge to commit suicide. So there's an urgent need to find out those words and inform the concerned authorities. There are many detectors to detect if the posts are suicidal. Some researchers made Speech-based Suicidal Ideation Detection [8] and even made Depression and Suicide Risk Detectors From Internet Usage Traces [9] but these are mostly in English. However, the study shows that when people tend to be emotional, they speak and post in their mother tongue [10]. The large population of around 290 million Bengali speakers [11] and a suicide rate of 7.3% in Bangladesh alone [12], where according to a Dhaka Tribune study [13], 513 student suicides in the 2023 year, who are in that particular age group of suicide makes it obvious that they may post about their feeling on social media

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

in their native language before doing any unethical task like taking their lives. These statements create a dire need for a detector that can verify the mental states of them and their friends easily in their native language so that if there's some inconvenience found, help can be sent to them.

To save natives' lives by detecting their thoughts indicating any mental instability and provoking suicide, the contribution of the paper is as follows:

1) Creation of a customized dataset of Bangla posts containing 2590 data, collected from social media platforms, particularly Facebook and Twitter.
2) Comparing the performance of advanced models e.g. BiLSTM, BERT, and XLM-RoBERTa on the customized dataset and choosing the BERT model for an accuracy of 92%.
3) Construction of a Chrome extension which if any text is highlighted in social media will predict if there's any suicidal thought accurately.

II. Literature Review

Considering the necessity of addressing the issue and saving individuals from fatal consequences, a lot of research has been conducted to detect mental health deterioration and predict suicidal tendencies from the social media activities of users. Although adequate works are available in English, only a few can be found in low-resource languages like Bangla. In [14] Islam et al. developed a suicidal attempt prediction system from social media posts creating a dataset called BanglaSPD. They compared various machine learning and deep learning models trained on this dataset, including logistic regression, SVM, CNN, LSTM, and BiLSTM. Their best-performing model was a CNN+BiLSTM architecture using FastText word embeddings, which achieved an F1 score of 0.61. However, their models are backdated compared to state-of-the-art systems with advanced transformer models yielding a higher accuracy [15]. Mohammed et al. [16] proposed a similar system, where they applied an ensemble approach for depression analysis from social media data in Bangla. The authors used a modified feature selection method combining TF-IDF, Extra Tree Classifier, and Principal Component Analysis, achieving 92.80% accuracy with eXtreme Gradient Boost (XGB). Our study 'Suicide Guard' uses more advanced transformer models e.g. BERT, XLM-RoBERTa, and offers practical application by developing a Chrome extension for suicidal thought prediction. A few systems have also been found using transformer models along with the use of machine learning, and deep learning for Bengali depressive text classification [17]. Although the authors claimed XLM-R as their best-performing model with an accuracy of 60.89%, there's still room for improvement, which has been achieved by 'Suicide Guard'. A greater accuracy has been achieved by Khan et al. [18] with recurrent neural networks and long short-term memory algorithms, yet the system is only limited to classifying Bangla social media posts as 'happy' or 'sad'. 'Suicide Guard' offers a simple browser extension application in addition to using sophisticated transformer models for prediction.

Besides suicidal thought detection from low-resource languages like Bangla, several works in English posts have been found and analyzed to identify gaps and improve our proposed system. Haque et al. [19], compared the performance of several machine learning (ML) and deep learning (DL) models in detecting suicidal ideation in tweets and found Bi-LSTM as the most effective one. The authors pointed out the necessity of a web application for real-time identification of suicidal thoughts, whereas 'Suicide Guard' comes equipped with such a feature while developed on a more efficient model. Other than neural networks, ensemble models such as Random Forest have also been used for suicidal ideation detection from tweeter posts [20]. Despite satisfactory accuracy, these models lack appropriate implementation to help prevent suicidal acts. Apart from this, transformer models and transfer learnings are being adopted as one of the most effective ways for text analysis to predict suicidal thoughts. In [21] and [22] the authors compared several deep learning and transformer models to evaluate their performance on suicidal thought detection and identified RoBERTa as the best-performing model. 'Suicide Guard' leverages the extended version of RoBERTa, the XLM-RoBERTa, that obtained satisfactory accuracy on the customized Bangla social media post dataset.

III. Methodology

The implementation of the system follows a structured approach from data collection, preprocessing, and model training to embedding the machine learning model into the Chrome extension for ultimate prediction. Figure 1 presents the steps followed for the system development.

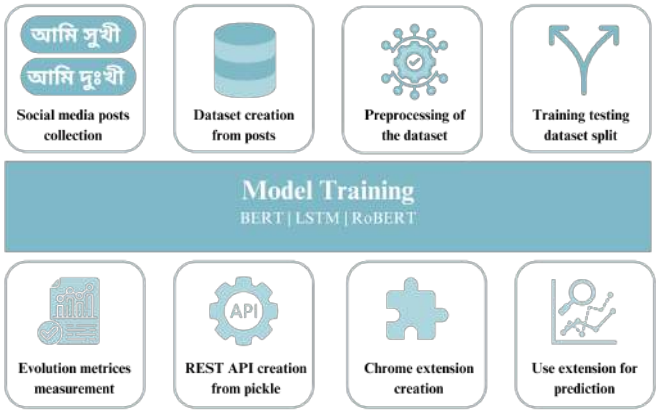

Fig. 1. Step-by-Step Approach to the System Development

A. Dataset Creation

The primary and crucial part of a data-driven model is the collection of relevant data from authentic sources. For this study, a primary source of data collection has been identified as Facebook and Twitter. Sentimental posts have

been extracted and curated from these platforms to create a uniformly balanced dataset. The final dataset is then formed with 2590 data which has been uploaded to the HuggingFace repository. Dataset link: Bangla-Suicidal-Post-Dataset

B. Preprocessing

The preprocessing steps vary to some extent from model to model. Nonetheless, some commonalities are still present in all of them.

* **Text Normalization:** Removing punctuation, converting to lowercase, and normalizing Unicode characters.
* **Tokenization:** Splitting texts into tokens. Different algorithms have their dedicated tokenizers for instance, BERT's WordPiece tokenizer.

Distinct preprocessing steps for LSTM:

- **Vocabulary Creation:** Building a vocabulary of unique words from the tokenized texts. A unique integer ID is assigned to each word of the vocabulary.
- **Truncation and Padding:** Ensuring all sequences are of uniform length.
- **Embedding:** Mapping each word in the vocabulary to its corresponding embedding vector.

Distinct preprocessing steps for BERT and XLM-RoBERTa:

- **Special Tokens Addition:** Adding special tokens such as [CLS] at the beginning of input for classification tasks, [SEP] to separate segments in the class e.g. question answers, etc.
- **Attention Masking:** Creating attention masks to differentiate between real tokens and padding tokens.
- **Segment IDs:** In tasks involving sentence pairs, segment IDs (0 or 1) indicate which token belongs to which class.

C. Model Training

Training is the most substantial part of model development. This key segment needs to be carefully coded to obtain satisfactory accuracy. Each algorithm has different libraries and parameter settings for training their models. For the three different algorithms compared in this study, the approach is explained as follows-

BiLSTM

Bidirectional LSTM is a form of LSTM (Long-short term memory) that consists of two LSTM layers for input processing both from forward and backward directions. Each of these LSTM networks returns a probability vector as output and the final output is the combination of these probabilities.

$$p_t = p_f + p_b \tag{1}$$

Here,
p_t = Final probability vector
p_f = Probability vector from the forward network
p_b = Probability vector from the backward network

To develop the suicidal thought prediction system, the model settings are as follows: *model = sequential; number of dense layers = 9; activation function = softmax; optimizer = adam.* This model first initializes a sequential model and adds an embedding layer to convert word indices into dense vectors. It then adds a bidirectional LSTM layer with 512 units, followed by a Flatten layer to convert the 3D LSTM output into 2D. A dense layer with 9 units and a softmax activation is added for multi-class classification. The model is compiled with Adam optimizer and sparse categorical cross-entropy loss, then trained for 40 epochs with training and validation data.

BERT

BERT is the acronym for Bidirectional Encoder Representations from Transformers. It is pre-trained on two distinct NLP tasks:

1) Mask Language Model (MLM)
2) Next Sentence Prediction (NSP)

For the suicidal thought prediction, the model parameter settings are as follows: *Model = bert-base-multilingual-cased; optimizer = AdamW; epoch = 5.* This model follows sequential classification with k-fold cross-validation, setting the value of k=2. For each fold, it splits the data into training and validation sets and initializes the BERT model and optimizer. The model is trained over 5 epochs per fold, with training loss computed and gradients updated.

XLM-RoBERTa

RoBERTa is an extended version of BERT and stands for Robustly optimized BERT approach. It is the multilingual version of RoBERTa and is pre-trained in 100 different languages CommonCrawl filtered data of 2.5TB.

The parameter settings for this suicidal thought detection model with XLM-RoBERTa are as follows: *Model = FacebookAI/xlm-roberta-base; optimizer = AdamW; num_epochs = 8.* After splitting the data into train and validation sets, the model creates dataloader objects and initializes the AdamW optimizer with a learning rate of $2e-5$. It then enters a training loop with 8 epochs. For each batch in the training DataLoader, it zeros the gradients, computes the model's outputs and loss, performs backpropagation, and updates the model's parameters. Calculates the average training loss at the end of each epoch.

D. Evaluation Metrics

Four widely acceptable evaluation metrics were determined for each of these models.

Accuracy: It is defined as the percentage of accurate predictions with respect to all the predictions made.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

Here,
TP = (True Positive) Model accurately predicting positive data

TN = (True Negative) Model accurately predicting negative data
FP = (False Negative) Model wrongly predicting negative data
FN = (False Negative) Model wrongly predicting positive data

Precision: This metric is calculated as a ratio of true positive predictions with all the positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall: This is also termed sensitivity and is a measure of the number of times a model correctly identifies positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-Score: This is the measure of the harmonic mean of precision and recall.

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{5}$$

E. Rest API Creation

After choosing the right algorithm which is BERT, a pickle file was created from the model. Pickling of the model is the process of converting Python objects into byte streams for transporting data over the network. The pickle file is mainly downloaded to make REST API using the FLASK framework. Flask is a Python-made web framework which is well-suited for making REST API. REST is an Application Programming Interface(API), as the API's task is to enable communication between different systems. REST also enables communication using HTTP requests. In Figure 2, it can be seen that when text

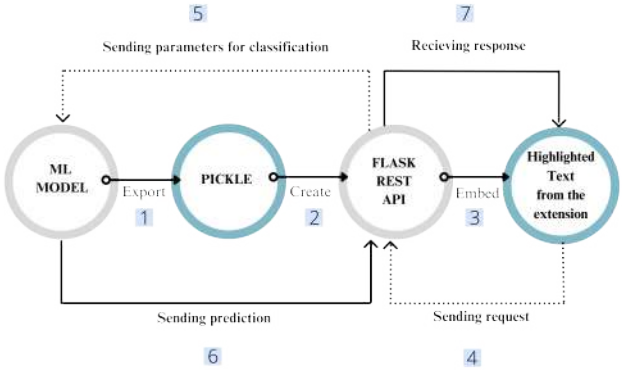

Fig. 2. ML Model Embedding Workflow

is highlighted the Chrome extension sends it to Flask REST API which sends an HTTP request to the model which is already uploaded in the Google Cloud. The model receives requests with text as a parameter after that model predicts based on the text which will then be sent to Flask REST API which then gives prediction as a response to the Chrome extension which will be shown in the extension if the text has suicidal thoughts or not.

F. Chrome Extension Creation

For creating a Chrome extension the (manifest.json) file was created. This provides metadata about the extension which also defines permissions, configuration, and capabilities of the extension. This JSON uses a (popup.html) file. This HTML is the actual design of the Chrome extension so the button and header have been defined here along with the (popup.js) file. This javascript file requests via POST method to API for prediction by sending 'highlighted text' as the element. Then an icon needs to be shown in the extension tab/bar so a picture in these sizes should be taken: 16*16 pixel, 48*48 pixel, and 128*128 pixel. Lastly, a (background.js) file to do some background tasks.

G. Connecting Chrome Extension to REST API

Figure 3 illustrates the process. Firstly, a zip folder of all the files (manifest.json, popup.html, background.js, icon.png, and popup.js) was created. Secondly, the developer mode was turned on after going to the Chrome browsers (manage extensions) page. Then the extension package was uploaded by clicking on the Load unpacked and selecting the zip folder including all files. Figure 4 shows the loaded extension ready to be used.

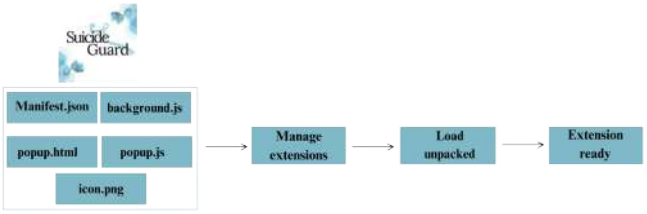

Fig. 3. Chrome Extension Creating Process Diagram

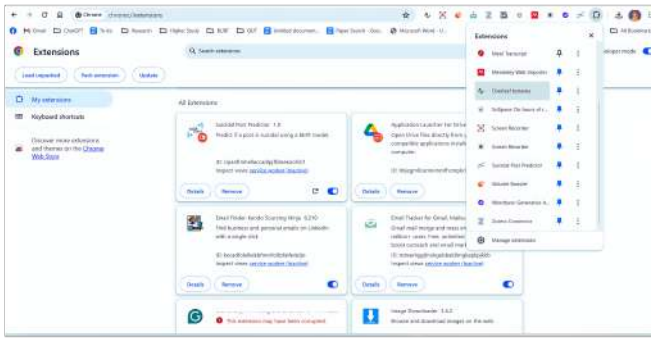

Fig. 4. Addition of the SuicideGuard Extension to Chrome Extension Manager

H. Use of Extension for Prediction

Now by highlighting or copying text into the extensions text box and clicking on the prediction button, the model will predict if one is suicidal or not as can be seen in Figure 5 and 6.

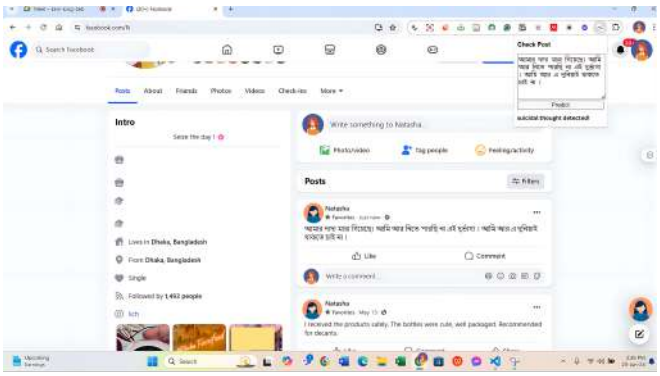

Fig. 5. Snapshot of Suicidal Post Identification from Facebook Interface I

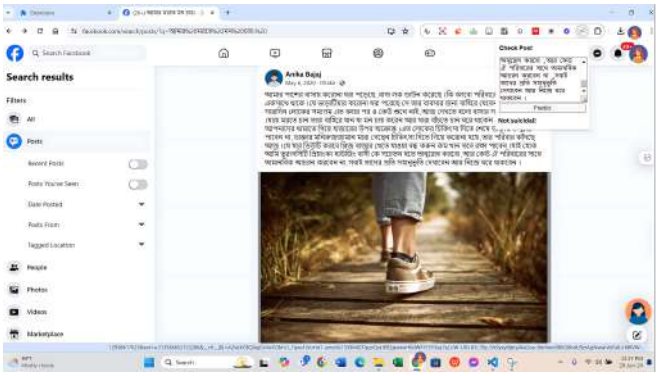

Fig. 6. Snapshot of Suicidal Post Identification from Facebook Interface II

IV. Result Analysis

Performance evaluation is an indispensable part of any machine learning-based system development. This section reflects the validity of the model in terms of its ability to accurately distinguish between true and false classes. For the 'Suicide Guard' system, the performance of BiLSTM, BERT, and XLM-RoBERTa has been evaluated to compare and select the best-performing model to be integrated into the predictive extension. Table IV shows the performance of the models in terms of accuracy, precision, recall, and f1-score. From the table, it is evident that the BERT model has the best performance on the dataset under study. For this reason, the BERT model has been further analysed and chosen for deployment in the Chrome extension. Figure 7 depicts the confusion matrix of the BERT model.

TABLE I
Performance Evaluation of the Machine Learning Models

	BiLSTM	BERT	XLM-RoBERTa
Accuracy	0.78	0.92	0.88
Precision	0.7804	0.9204	0.8804
Recall	0.78	0.92	0.88
F1-score	0.78	0.92	0.88

After machine learning model development and selection, a pickle file of the model has been extracted to integrate it with the Chrome extension. This task required the use of API (Application Programming Interface) that enables the communication between the machine learning model at the backend and the Chrome extension interface at the frontend. The Postman API platform has been used to perform the crucial task. Figure 8 and 9 exhibit the outcome of prediction from the postman interface.

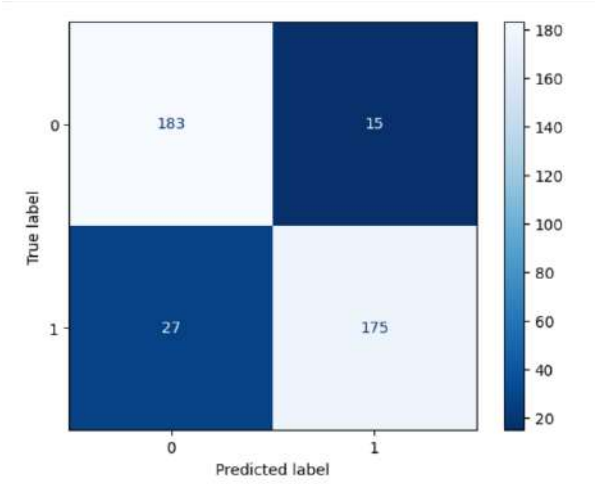

Fig. 7. Confusion Matrix of the BERT Model

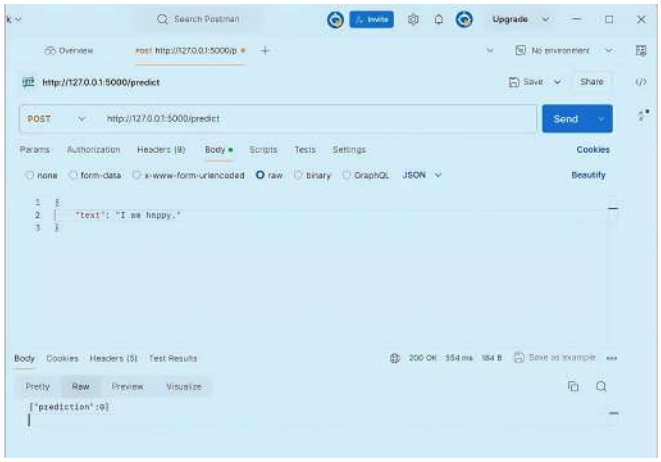

Fig. 8. Snapshot of Prediction from Postman API I

V. Conclusion and Future Works

SuicideGuard is a pioneering Chrome extension that uses the BERT model to identify suicidal thoughts. It's unique from other systems as it is based on Bangla posts focusing on Bengali speakers and the Chrome extension which can predict in real-time as this situation demands a swift response. The paper presents a novel and helpful tool that can accurately predict suicidal thoughts by creating a customized dataset in Bangla, which has been achieved with the satisfactory performance of the transformer BERT model and the practical application of the Chrome extension. In the future, a counsellor chatbot using generative AI will be integrated to interact with

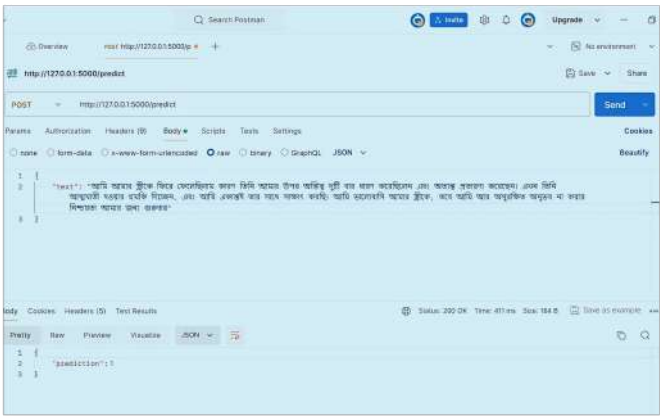

Fig. 9. Snapshot of Prediction from Postman API II

individuals who are having depressive thoughts which will create a wholesome system that can be used to help thousands and save indefinite lives.

REFERENCES

[1] World Health Organization: WHO, "Suicide," Aug. 28, 2023. https://www.who.int/news-room/fact-sheets/detail/suicide

[2] "Suicide rates in the US are on the rise: New study offers surprising," CU Boulder Today, Feb. 21, 2024. https://www.colorado.edu/today/2024/02/15/suicide-rates-us-are-rise-new-study-offers-surprising-reasons-why

[3] "Children and young people: statistics," Mental Health Foundation. https://www.mentalhealth.org.uk/explore-mental-health/statistics/children-young-people-statistics

[4] J. Howarth, "Top 35 social media platforms (June 2024)," Exploding Topics, May 30, 2024. https://explodingtopics.com/blog/top-social-media-platforms

[5] Statista, "Age distribution of internet users worldwide 2024," Statista, May 22, 2024. https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

[6] T. R. Soron and S. M. S. Islam, "Suicide on Facebook-the tales of unnoticed departure in Bangladesh," Global Mental Health, vol. 7, Jan. 2020, doi: 10.1017/gmh.2020.5

[7] J. Doom, "Amid onslaught of lawsuits, Instagram and Facebook will start hiding posts about suicide, self-harm and eating disorders from teenagers' accounts," Fortune, Jan. 10, 2024. https://fortune.com/2024/01/09/facebook-instagram-hide-posts-suicide-self-harm-eating-disorders-meta-lawsuits-zuckerberg/

[8] A. Pillai et al., "Investigating generalizability of speech-based suicidal ideation detection using mobile phones," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 7, no. 4, pp. 1–38, Dec. 2023, doi: 10.1145/3631452

[9] E. Marchese, "Depression and suicide risk detection from internet usage traces," Mar. 11, 2024. https://urn.fi/URN:NBN:fi:aalto-202403172759

[10] C. L. Harris, J. B. Gleason, and A. Ayçiçeği, "10. When is a First Language More Emotional? Psychophysiological Evidence from Bilingual Speakers," in Multilingual Matters eBooks, 2006, pp. 257–283. doi: 10.21832/9781853598746-012

[11] "Bengali - Worldwide distribution," Worlddata.info. https://www.worlddata.info/languages/bengali.php

[12] S. R. Mashreky, F. Rahman, and A. Rahman, "Suicide kills more than 10,000 people every year in Bangladesh," Archives of Suicide Research, vol. 17, no. 4, pp. 387–396, Oct. 2013, doi: 10.1080/13811118.2013.801809

[13] "Study: Bangladesh records 513 student suicides in 2023," dhakatribune.com. https://www.dhakatribune.com/bangladesh/337897/study-bangladesh-reports-513-student-suicides-in

[14] S. Islam, Md. S. A. Forhad, and H. Murad, BanglaSAPM: A deep learning model for suicidal attempt prediction using social media content in Bangla. 2022. doi: 10.1109/iccit57492.2022.10055237

[15] M. M. Rahman, Md. A. Pramanik, R. Sadik, M. Roy, and P. Chakraborty, Bangla documents classification using transformer based deep learning models. 2020. doi: 10.1109/sti50764.2020.9350394

[16] M. B. Mohammed, A. S. Md. Abir, L. Salsabil, M. Shahriar, and A. Fahmin, Depression analysis from Social Media Data in Bangla Language: An ensemble approach. 2021. doi: 10.1109/etcce54784.2021.9689887

[17] O. Adukwu, "Fault diagnosis in gas lift system using PDF data," Journal of Engineering Advancements, pp. 35–39, Jun. 2023, doi: 10.38032/jea.2023.02.001

[18] Md. R. H. Khan, U. S. Afroz, A. K. M. Masum, S. Abujar, and S. A. Hossain, "A Deep Learning Approach to Detect Depression from Bengali Text," in Advances in intelligent systems and computing, 2021, pp. 777–785. doi: 10.1007/978-981-33-4367-2_74

[19] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A comparative analysis on suicidal ideation detection using NLP, machine, and Deep learning," Technologies, vol. 10, no. 3, p. 57, Apr. 2022, doi: 10.3390/technologies10030057

[20] E. R. Kumar, K. V. S. N. R. Rao, S. R. Nayak, and R. Chandra, "Suicidal ideation prediction in twitter data using machine learning techniques," Journal of Interdisciplinary Mathematics/Journal of Interdisciplinary Mathematics, vol. 23, no. 1, pp. 117–125, Jan. 2020, doi: 10.1080/09720502.2020.1721674

[21] F. Haque, R. U. Nur, S. A. Jahan, Z. Mahmud, and F. M. Shah, A Transformer Based Approach To Detect Suicidal Ideation Using Pre-Trained Language Models. 2020. doi: 10.1109/iccit51783.2020.9392692

[22] T. K. Mishra, G. Sucharitha, N. Siddhartha, B. Raju, and S. N. Mohanty, Prediction of suicidal behaviour among the users on social media using NLP and ML. 2024. doi: 10.1109/esic60604.2024.10481588