Suicide Risk Assessment Using NLP And Machine Learning

1st Hazra Mohammed Ahnaf Faiyaz
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
hazra.mohammed.ahnaf.faiyaz@g.bracu.ac.bd

2nd Sadman Sakib Nabil
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

3rd Annajiat Alim Rasel
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

4th Md. Sabbir Hossain
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

5th Mehnaz Ara Fazal
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

***Abstract*—In contemporary society, the escalating incidence of suicide poses a significant and growing concern. Addressing this issue requires a profound understanding of its inherent risks and the development of effective strategies for risk reduction. This paper aims to contribute to this imperative by presenting a holistic approach to suicide risk assessment, utilizing the integration of Natural Language Processing and Machine Learning (ML) models.The primary focus of this study is to enhance the precision of suicide risk prediction through the meticulous analysis of textual data associated with at-risk individuals. A dataset "Suicide and Depression Detection" which is available on Kaggle has been utilized to unveil subtle linguistic patterns indicative of suicide risk. The textual data has been converted to lowercase, the punctuations have been removed to ensure clean text, and stemming has been performed to find the root words. TF-IDF Vectorizer has been used to vectorize the textual data. To discern the most accurate predictive model, various machine learning algorithms: Naive-Bayes Classifier, Random Forest, Decision Tree, Gradient Boosting, and K-Nearest Neighbor, were applied.**

I. Introduction

Suicide is a multidimensional and intricate social issue that cuts beyond national, cultural, and economic lines. The World Health Organization (WHO) estimates that suicide claims the lives of about 800000 people each year (1), making it a major cause of mortality worldwide. This widespread problem not only has a significant effect on people and their families, but it also presents significant obstacles for public health systems across the globe.

For preventative and intervention methods to be effective, it is vital to comprehend the complex elements that are associated with suicide. Suicidal thoughts and behaviors can be triggered by a variety of causes, including mental health issues, social stigma, economic inequality, and restricted access to mental health resources. Comprehensive research is therefore desperately needed in order to improve our comprehension of these variables and create focused strategies to lower suicide rates globally.

This paper focuses on the performance metrics of some common machine learning models along with NLP tasks to determine the scope of these techniques to further clarify and provide a comprehensible understanding of the applications of these techniques.

It starts off by performing NLP tasks on the "Suicide and Depression Detection" which is available on Kaggle, by converting the textual data to lowercase and removing the punctuations. It then performs stemming to secure the root words post which vectorization is conducted via the TF-IDF vectorizer. With the clean textual data that remains afterwards, some common machine learning algorithms are utilized to train and test models. The models are Naive-Bayes Classifier, Random Forest, Decision Tree, Gradient Boosting, and K-Nearest Neighbor out of which Naive-Bayes gave us the best results. This diverse set of models underwent rigorous evaluation to determine their efficacy in correctly predicting instances of suicide.The results are then shown on a historam containing the f1-scores of the aforementioned models. The Naive-Bayes Classifier model has an f1-score of 0.8819904 (Highest in our testing), Random Forest has 0.7655578, Decision Tree has 0.7699456, Gradient Boosting has 0.7444396 (Lowest in

our testing), and K-Nearest Neighbor has 0.8497434. Further processing and a larger, fine-tuned, and unique dataset will result in better outcomes.

II. Literature Review

Suicide has become a prominent and concerning issue in the 21st century, with rates doubling compared to the previous century. Extensive research indicates that mental illness plays a significant role in most suicide cases, often posing challenges in providing concrete evidence for such incidents due to their psychological nature (Hogan and Grumet). Global responses to rising suicide rates have led to increased awareness and initiatives within various organizations. Strategies are being implemented to mitigate suicide rates by enhancing public understanding of mental health issues and ensuring effective treatment options according to the United States Surgeon General and National Action Alliance for Suicide Prevention, New Zealand Ministry of Health and NHS England.

According to United States Surgeon General and National Action Alliance for Suicide Prevention, a noteworthy concept and policy in suicide prevention is the emergence of "Zero Suicide," emphasizing the paramount importance of preventing suicides within healthcare organizations . Studies suggest that a decline in suicide rates can be achieved through proactive mental health care. Individuals grappling with psychological health issues are found to be more prone to suicide attempts. While Hogan doesn't explicitly state that mental health is the predominant factor in suicide cases, it remains a research gap. Addressing this gap involves providing comprehensive guidance on mental health to hospitals, families, and individuals dealing with psychological health issues, thereby contributing to suicide prevention (Mulder, Newton – Howes, and Coid 2016).

Kant, in his work "Groundwork of the Metaphysics of Morals," argues that individuals contemplating suicide may see it as a means to end their suffering rather than an intentional act of self-harm (Kant 2002). He emphasizes the responsibility of each individual to preserve their life rather than surrendering to despair. Consequently, the ultimate challenge in preventing suicides lies in the care of mental health. In 2023, NLP strategies are increasingly becoming an important mechanism to sort textual data into more concise, comprehensible, and clean data. According to 'The Recent Advances and Applications of Natural Language Processing' by Bhoi, P. C., Singh, D., and Das, B. R. (May 2023) they mention some trends in 2023 related to NLP, such as virtual assistants, text summarization, sentiment analysis etc.

III. Methodology

We used a dataset "Suicide and Depression Detection" available on Kaggle to explore the capabilities of Natural Language Processing and some common Machine Language models. Posts from the Reddit platform's "SuicideWatch" and "depression" subreddits are included in the dataset. The Pushshift API was used to gather these posts. The dataset comprised posts from "SuicideWatch" between December 16, 2008, when it first started, and January 2, 2021. Posts from January 1, 2009, to January 2, 2021, were gathered under the category "depression." The dataset consists of approximately 232074 texts from the aforementioned subreddits out of which we have used 10000 sample texts and split it into two parts: Suicide - 4952 and Non-suicide - 5048 to balance the outcomes. In our testing, we incorporated samples of the texts of 1000, 5000, 10000, 20000, 50000, and 100000. We noticed that 10000 sample texts generate an accuracy of 0.88 and do not increase significantly in case more than 10000 samples are used. The paper starts off with loading the dataset into the code. The dataset has a shape of 232074 and 3 columns of which one is an unnamed column, one is text column, and the last one is the class column. The unnamed column contains the number of the text in the whole dataset, the text column contains all the textual data retrieved from the aforementioned subreddits, and the class column contains the binary classification values (Suicide and Non-suicide). Firstly, we changed the textual data into lower cases and removed all the punctuations to clean the data for further processing.We then tokenized and removed stopwords, performed stemming to find the root of the word, and then vectorized the data using TF-IDF vectorizer. After these steps were completed, we incorporated the machine learning models (Naive-Bayes Classifier, Random Forest, Decision Tree, Gradient Boosting, and K-Nearest Neighbor) to determine which of these models perform the best in identifying the instances of suicide and non-suicide. We also included code that incorporates the best performing model, takes an input, and then predicts if it is a suicide or a non-suicide.

IV. Result Analysis

Dummy

V. Discussion

Dummy

VI. Future Work and Limitations

Dummy

VII. Conclusion

Dummy

A. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

Fig. 1. Example of a figure caption.

TABLE I
TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	*Subhead*	*Subhead*
copy	More table copy[a]		

[a]Sample of a Table footnote.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] World Health Organization. (n.d.). Suicide. Retrieved December 9, 2023, from https://www.who.int/news-room/fact-sheets/detail/suicide

[2] Bhoi, P. C., Singh, D., Das, B. R. (May 2023). The Recent Advances and Applications of Natural Language Processing. Paper presented at the 2nd NCRTCEA 2003, SOCSE, Sandip University, Nashik, India. Institute of Technical Education and Research; Siksha O Anusandhan University; NM Institute of Engineering Technology Bhubaneswar.

[3] Fernandes, A.C., Dutta, R., Velupillai, S. et al. Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. Sci Rep 8, 7426 (2018). https://doi.org/10.1038/s41598-018-25773-2.

[4] Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide Note Classification Using Natural Language Processing: A Content Analysis. Biomedical Informatics Insights. 2010;3. doi:10.4137/BII.S4706.

[5] Jain, P., Srinivas, K. R., Vichare, A. (2022). Depression and Suicide Analysis Using Machine Learning and NLP. Journal of Physics: Conference Series, 2161, 012034. Published under license by IOP Publishing Ltd. doi: 10.1088/1742-6596/2161/1/012034.

[6] Dummy.

[7] Dummy.