Comparative Analysis of Machine Learning Algorithms for Detecting Suicidal Ideation in Social Media Content

Jabez Ian Chris D. Peñalver, Cristopher C. Abalorio
Caraga State University - Main Campus, Butuan City, Philippines
jdpenalver@carsu.edu.ph

Abstract— Suicidal ideation in students is one of the most serious threats that academic institutions have to deal with, as studies have shown that 10-24% of college/university students have thought of taking their own life seriously. Presently, all current intervention strategies practiced in campus guidance offices tend to be reactive crisis management only and do not seem to focus on preventing the emergence of suicidal thoughts. This limitation is addressed in the present study by proposing a Suicide Ideation Analysis and Monitoring (SIAM) System that is grounded on the use of Artificial Intelligence and Natural Language Processing to assess suicidal ideation in social media content at an early stage. The research work made use of a large dataset containing 232,074 posts in total, obtained from a Reddit-based community on SuicideWatch. Seven machine learning algorithms were applied together with 10-fold cross validation: Random Forest, Naive Bayes, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor, Logistic Regression and Gradient Boosting. The models were evaluated using several performance metrics, including accuracy, precision, recall and F1 score. Overall, the SVM classifier was the best performing one as it produced 93% accuracy for every other metric, compared to the 90% accuracy of Random Forest and 92% accuracy of Logistic Regression. Also, support for Anthropic AI system provides an option for tagging and translating Filipino (Tagalog) and Cebuano to English which enables its possible use in schools and universities in the Philippines. The research demonstrates the potential of AI-driven solutions in transforming suicide prevention efforts from reactive to proactive approaches, while maintaining privacy considerations through careful handling of social media data. The developed system represents a significant advancement in campus mental health support systems, providing automated, privacy-conscious monitoring of suicide risk indicators that could enable earlier intervention and support for at-risk students.

Keywords – Suicide Ideation Detection, Machine Learning Classification, Natural Language Processing, Student Mental Health

1 Introduction

Suicide remains a pressing global public health issue, claiming over 703,000 lives annually and ranking as the second leading cause of death among individuals aged 15–24 worldwide in 2023 according to WHO. Suicidal ideation, a precursor to attempts, is alarmingly prevalent among students, with 10–24% of college and university students reportedly considering suicide [1] . Key risk factors include mental health disorders (e.g., depression, anxiety), substance abuse, academic and financial stress, and social isolation

[2]. Vulnerable populations, such as LGBTQ+ students, face elevated risks due to discrimination [3], while recent events like the COVID-19 pandemic have intensified mental health challenges among students [4].

Current campus interventions for suicidal ideation are predominantly reactive. Guidance offices typically respond post-crisis by interviewing affected students, consulting peers or instructors, and reviewing recent social media activity [5][6][7]. This approach lacks proactive measures to identify and support at-risk students preemptively, highlighting a critical gap in prevention strategies.

Artificial intelligence (AI), particularly natural language processing (NLP), offers a promising solution for early detection. Chiroma et al. [8] demonstrated NLP's ability to classify suicide-related tweets, while Aldhyani et al. [9] trained models on social media datasets to identify linguistic patterns linked to suicidal ideation. Fonseka et al. [10] further emphasize predictive analytics' potential to enable timely interventions, such as emotional support and crisis alerts. Building on these findings, this study proposes the Suicide Ideation Analysis and Monitoring (SIAM) system, leveraging NLP and binary classification models (e.g., Random Forest, Naive Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Gradient Boosting) to detect suicidal ideation in student populations proactively. By analyzing text data, such as social media posts, SIAM aims to facilitate upstream prevention, addressing the limitations of reactive campus strategies and reducing the escalation of suicidal tendencies among students.

2 Related Literature

This section reviews existing literature and studies pertinent to the research, encompassing suicide and suicidal ideation, data collection, text preprocessing, and classification models for detecting suicidal and non-suicidal text. The review draws from diverse sources, including journals, books, websites, theses, and dissertations, to establish a foundation for the proposed study.

2.1 Depression, Suicide, Suicidal Ideation, and Artificial Intelligence

Depression, a prevalent mental health condition, varies in severity and is categorized as mild, moderate, or severe [11] . Moderate to severe cases often correlate with increased risks of self-harm, suicidal ideation, or suicide attempts. Janota et al. [6] underscore the necessity of professional intervention when suicidal thoughts emerge or behavioral shifts suggest suicidal tendencies, emphasizing timely support as critical. Similarly, Mofatteh [2] advocates a multifaceted suicide prevention strategy addressing mental health, support systems, and resilience. Suicidal ideation, a precursor to attempts, demands recognition and compassionate intervention to avert escalation [2].

Artificial intelligence (AI) has emerged as a transformative tool in suicide prevention. Bernert et al. [12]highlight AI's capacity to analyze large datasets for risk detection, addressing the global challenge of predicting suicide. Fonseka et al. [10] note that predictive analytics can identify individuals in crisis, enabling targeted interventions, while population-level algorithms pinpoint at-risk groups, informing policy and resource allocation. AI also supports clinical management, offering efficient, adaptable solutions for diagnostics and therapy, particularly in underserved areas [10].

2.2 Data Collection and Labeling

Effective suicidal ideation detection relies on quality datasets. Aldhyani et al. [9] utilized Reddit data from the "SuicideWatch" and "depression" subreddits, collected via the Pushshift API from 2008 to 2021. The dataset, comprising 232,074 posts (116,037 suicidal, 116,037 non-suicidal), was labeled based on subreddit origin, with non-suicidal posts sourced from "r/teenagers" [9]. This approach demonstrates the potential of social media data for training AI systems like the proposed Suicide Ideation Analysis and Monitoring (SIAM) system.

2.3 Text Preprocessing

Text preprocessing is critical for preparing data for classification. Chiroma et al. [8] and Jain et al. [13] employed natural language processing (NLP) techniques, including punctuation removal, lowercasing, tokenization, stop word removal, and lemmatization/stemming. These steps enhance data quality by reducing noise and dimensionality, ensuring compatibility with machine learning models [8][13].

2.4 Feature Extraction

Feature extraction transforms text into numerical representations for analysis. Aldhyani et al. [9] applied Term Frequency-Inverse Document Frequency (TF-IDF), which weights words based on their frequency in a document relative to the corpus [14]. TF measures term occurrence within a document, while IDF reduces the influence of common terms, prioritizing discriminative features [14]. This method effectively captures relevant text characteristics for classification tasks [9].

2.5 Binary Classification Studies

Binary classification of suicidal text has been widely explored. Khatun et al. [15] evaluated Naive Bayes, Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) on a 5,572-SMS spam dataset, achieving varying accuracy levels. Chiroma et al. [8] classified 2,000 tweets using Decision Trees, Naive Bayes, Random Forest, and SVM, achieving a 0.779 accuracy with Decision Trees, despite challenges posed by Twitter's brevity and noise [8]. These studies inform algorithm selection for suicidal text detection.

2.6 Classification Evaluation Metrics

Evaluation metrics assess model performance comprehensively. Agrawal [16] emphasizes the confusion matrix, detailing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), enabling metrics like accuracy, precision, recall, and F1-score [16]. Accuracy measures overall correctness, precision evaluates positive prediction reliability, recall assesses positive case detection, and F1-score balances precision and recall [16][17]. These metrics ensure robust model validation beyond mere accuracy [17].

3 Methods

This section outlines the systematic approach and methods employed in conducting this research. It details the research design, the techniques involved for data collection and preprocessing, and the development processes that was implemented. The methodology serves as a comprehensive blueprint, ensuring the study's validity, reliability, and adherence to ethical standards.

3.1. Dataset

This phase was split into multiple stages: first, the dataset was divided into training and validation sets for model development and k-fold cross-validation. The best performing model was then saved and used to classify real-world social media data to evaluate its practical performance.

Training Data Set. This research made use of the publicly accessible Reddit data obtained from the Kaggle website. This dataset was previously employed in the study conducted by Aldhyani et al. [9]. The dataset consisted of 232,074 posts from the SuicideWatch subreddit, spanning the period from December 16, 2008, to January 2, 2021. These posts were evenly divided between 116,037 posts as suicidal and 116,037 posts as non-suicidal. The posts were gathered using the Pushshift API [9].

3.2. Text Preprocessing and Feature Extraction

Text Preprocessing. The researchers preprocessed the textual posts by removing noise and irrelevant elements. It involved stop word removal, punctuation removal, lowercasing, tokenization, and lemmatization. The Natural Language Toolkit (NLTK) was employed to perform these basic preprocessing tasks on the dataset.

Feature Extraction. This research utilized Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF, a statistical method that weighs the importance of words in a document based on their frequency within the document and across the entire corpus.

3.3. Cross Validation

k-Fold Cross Validation. The procedure involves splitting the data sample into k groups. For each group, the group is treated as a test set while the remaining groups are used as a training set. A model is trained on the training set and evaluated on the test set. This process is repeated for each group, and the evaluation scores are retained. The skill of the model is then summarized based on the sample of evaluation scores. Cross-validation helps estimate the model's performance on unseen data and provides a less biased estimate compared to a simple train/test split. The procedure ensures that each observation in the data sample is used as part of the test set once and as part of the training set k-1 times.

For this research, k=10 was chosen for a dataset of 232,074 points, each fold would have approximately 23,207 data points (232,074 / 10 = 23,207.4). In each iteration, one-fold with 23,207 points would be used as the test set, and the remaining 9 folds (with a total of 208,867 points) would be used for training the model.

3.4. Evaluation Metrics

To evaluate the performance of the models in classifying post content as either suicidal or non-suicidal. It was evaluated using common binary classification evaluation matrices. The evaluation metrics that were employed are Accuracy, Precision, Recall, and F1-score, which will be calculated as shown in below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP \ + \ FN} \tag{3}$$

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision \ + \ Recall} \tag{4}$$

(TP = True Positive, FP = False Positive, FN = False Negative, TN = True Negative)

3.5. Model Training and Deployment

This study utilized, the machine learning algorithms specifically Random Forest, Naive Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression and Gradient Boosting; to develop the binary classification model. The model with the highest performance based on accuracy, precision, recall and f1-score was then saved as a pickle file, which was a serialized representation of the trained model that could be easily loaded and used for making predictions. The pickle file of the binary classification model with the highest performance was integrated into a Python Flask web application. This Flask backend was responsible for receiving text data from the mobile application, passing it through the loaded model for classification, and saved the predicted suicidal post to the database. Consequently, it was utilized for web-based analytics dashboard.

4 Results & Discussion

The study utilized a comprehensive dataset from Reddit's SuicideWatch subreddit, accessed through Kaggle. The complete dataset encompassed 232,074 posts collected between December 16, 2008, and January 2, 2021, with an equal distribution of 116,037 suicidal and 116,037 non-suicidal posts as shown in Figure 1.

Fig. 1. Class Distribution

The research implemented a robust cross-validation approach with k=10 to evaluate the performance of seven machine learning algorithms: Random Forest, Naive Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Gradient Boosting. Each model was evaluated using multiple performance metrics: Accuracy, Precision, Recall and F1 Score. The Table 1 that follows shows the results of evaluation:

Table 1. Evaluation Matrices Result

Classifier	Accuracy	Precision	Recall	F1 Score
Random Forest	0.90	0.90	0.90	0.90
Naïve Bayes	0.87	0.88	0.87	0.87
Decision Tree	0.84	0.84	0.84	0.84
Support Vector Machine	0.93	0.93	0.93	0.93
K-Nearest Neighbor	0.51	0.61	0.51	0.36
Logistic Regression	0.92	0.92	0.92	0.92
Gradient Boosting	0.88	0.88	0.88	0.88

The result shows that Support Vector Machine (SVM) emerged as the best-performing model with an accuracy of 0.93. Based on its superior performance, the SVM model was selected for deployment. SVM likely outperforms other models in suicide ideation detection due to its strength in high-dimensional, sparse data, typical of TF-IDF text representations. Its capacity to create clear decision boundaries maximizes the separation between suicidal and non-suicidal classes, essential for binary classification. Additionally, SVM's resilience to class imbalance and ability to capture subtle linguistic differences make it well-suited for identifying nuanced signs of ideation without overfitting [18]. This results in balanced performance across accuracy, precision, recall, and F1 score, which is critical for a sensitive task like suicide ideation detection.

Therefore, SVM works well for detecting suicidal thoughts in text because it can clearly separate different patterns of language used by people with suicidal and non-

suicidal thoughts. It handles complex text data effectively and can identify even small differences in wording or tone, which are often important in this context. SVM is also good at managing situations where one type of text is more common than the other (like non-suicidal thoughts). This helps avoid errors and ensures its reliability at identifying suicidal expressions accurately, which is crucial for this sensitive task.

The study utilized the Support Vector Machine (SVM) model, saved as a pickle file, to power the web analytics dashboard. A prototype mobile app was developed to collect students' basic information and social media posts, requiring students to input their ID numbers during login (Login via Facebook functionality) for proper tracking in the dashboard. The system uses Anthropic AI to detect and translate Cebuano or Filipino posts to English, while English posts are directly processed. For privacy, the Flask web application only stores analytics data of potentially concerning posts without keeping the actual content. The analytics and monitoring system updates whenever students log in through Facebook, suggesting its integration with the university's existing student portal would be beneficial for more consistent monitoring.

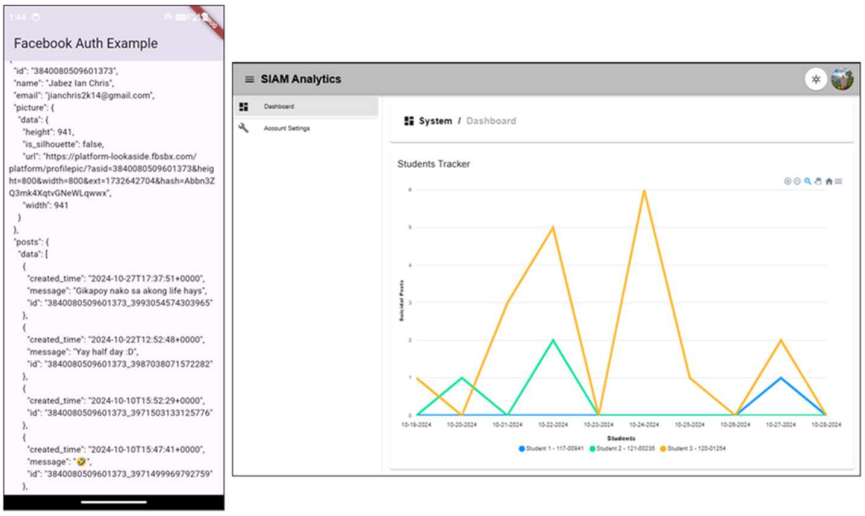

Fig. 2. Prototype Mobile App & Web Dashboard

The Figure 2 shows the successful integration of Facebook authentication to monitor student social media activity through a secure mobile app and web analytics system. The tracking dashboard revealed varying activity levels among three students over a 10-day period in October 2024, with Student 3 showing the highest activity peak of 6 suicidal posts, while Students 1 and 2 maintained lower levels around 1-2 suicidal posts. The system effectively processes posts in multiple languages (English, Cebuano, and Filipino) and automatically translates them as needed, demonstrating its capability to comprehensively monitor student expressions regardless of language preference.

5 Conclusion

This work demonstrates the efficacy of machine learning techniques for detecting and monitoring suicidal ideation through social media content analysis. A suite of classification models was developed and assessed, with the Support Vector Machine (SVM) achieving the highest performance, recording a 93% accuracy across precision, recall, and F1-score metrics. Key contributions include the validation of natural language processing (NLP) and machine learning for proactive identification of suicidal ideation, the superior performance of SVM over alternatives (e.g., Random Forest at 90% and Logistic Regression at 92%), the utility of TF-IDF feature extraction with comprehensive text preprocessing, and the integration of a multilingual framework leveraging Anthropic AI to translate Filipino (Tagalog) and Cebuano content into English.

The proposed system offers practical implications for suicide prevention in academic settings, particularly in the Caraga region, where 13.5% of youth aged 15-24 report suicidal contemplation. By enabling early detection of at-risk students, this approach shifts from reactive to proactive strategies, enhancing campus mental health support systems. Privacy considerations were addressed through careful data handling, ensuring compliance with ethical standards.

This research bridges a gap in automated mental health monitoring by providing a scalable, privacy-conscious tool for campus guidance offices. Future efforts may extend the dataset to encompass broader linguistic and cultural diversity, incorporate real-time monitoring, refine intervention protocols triggered by system alerts, and assess long-term impacts on suicide prevention outcomes. These findings contribute to the advancement of AI-driven interventions in mental health, underscoring their potential to address pressing public health challenges in educational environments.

References

[1] WHO, "WHO EMRO _ Suicide _ Health topics," *Suicide*, 2023.

[2] M. Mofatteh *et al.*, "Suicidal ideation and attempts in brain tumor patients and survivors: A systematic review," 2023. doi: 10.1093/noajnl/vdad058.

[3] E. K. Gill and M. T. McQuillan, "LGBTQ+ Students' Peer Victimization and Mental Health before and during the COVID-19 Pandemic," *Int J Environ Res Public Health*, vol. 19, no. 18, 2022, doi: 10.3390/ijerph191811537.

[4] N. Barberis, M. Cannavò, F. Cuzzocrea, and V. Verrastro, "SUICIDAL BEHAVIOURS DURING COVID-19 PANDEMIC: A REVIEW," *Clin Neuropsychiatry*, vol. 19, no. 2, 2022, doi: 10.36131/cnfioritieditore20220202.

[5] R. V. Lopiga, "Suicide Potential and Depression: Risk and Protective Factors among College Students in the Philippines," *The Normal Lights*, vol. 15, no. 2, 2021, doi: 10.56278/tnl.v15i2.1860.

[6] M. Janota, V. Kovess-Masfety, C. Gobin-Bourdet, and M. M. Husky, "Use of mental health services and perceived barriers to access services among college students with suicidal ideation," *J Behav Cogn Ther*, vol. 32, no. 3, 2022, doi: 10.1016/j.jbct.2022.02.003.

[7] A. M. Memon, S. G. Sharma, S. S. Mohite, and S. Jain, "The role of online social networking on deliberate self-harm and suicidality in adolescents: A

systematized review of literature," 2018. doi: 10.4103/psychiatry.IndianJPsychiatry_414_17.

[8] F. Chiroma, H. Liu, and M. Cocea, "Text Classification for Suicide Related Tweets," in *Proceedings - International Conference on Machine Learning and Cybernetics*, 2018. doi: 10.1109/ICMLC.2018.8527039.

[9] T. H. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. T. Ahmed, "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models," *Int J Environ Res Public Health*, vol. 19, no. 19, 2022, doi: 10.3390/ijerph191912635.

[10] T. M. Fonseka, V. Bhat, and S. H. Kennedy, "The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors," 2019. doi: 10.1177/0004867419864428.

[11] S. Chao, "Overview of Depression," 2024. doi: 10.1016/j.emc.2023.06.013.

[12] R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, N. H. Shah, and F. Abnousi, "Artificial intelligence and suicide prevention: A systematic review of machine learning investigations," 2020. doi: 10.3390/ijerph17165929.

[13] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and Suicide Analysis Using Machine Learning and NLP," in *Journal of Physics: Conference Series*, 2022. doi: 10.1088/1742-6596/2161/1/012034.

[14] C. Goyal, "Syntactic Analysis | Guide to Master Natural Language Processing(Part 11)," www.analyticsvidhya.

[15] A. Khatun, M. H. Matin, A. Miah, and R. Miah, "Comparative Study on Text Classification," *International Journal of Engineering Science Invention (IJESI)*, vol. 9, no. 9, 2020.

[16] S. K. Agrawal, "Metrics to Evaluate your Classification Model to take the right decisions," Data Science Blogathon.

[17] P. Vickers, L. Barrault, E. Monti, and N. Aletras, "We Need to Talk About Classification Evaluation Metrics in NLP," 2024. doi: 10.18653/v1/2023.ijcnlp-main.33.

[18] JavaTpoint, "Support Vector Machine Algorithm," JavaTpoint. [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm