Performance Evaluation of Deep Learning Models on Suicide Ideation Detection of Reddit Posts

Linda Meng

Received August 01, 2024
Accepted November 30, 2024
Electronic access December 15, 2024

This paper focuses on the issue of applying machine learning techniques to detect suicidal ideation within social media posts on Reddit. Comparisons of the models include a baseline decision tree classifier, a hybrid Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) model, and a transfer learning approach using the Bidirectional Encoder Representations from Transformers (BERT) language model. The research utilizes a dataset of 16,000 posts from r/SuicideWatch and r/teenagers subreddits, employing natural language processing and deep learning techniques to classify posts as suicide-related or non-suicide-related. Of the three models, The BERT model demonstrated the best performance, achieving an AUC-ROC of 0.9909 and an accuracy of 97%, outperforming both the LSTM-CNN (AUC: 0.9611, Accuracy: 90%) and Decision Tree (AUC: 0.9477, Accuracy: 88.16%) models. Feature importance analysis revealed key linguistic markers for suicide ideation, including not only explicit indicators of distress but also subtle contextual cues. Though these results show some promises of machine learning for the detection of suicide risk, they also demonstrate challenges in fine-tuning pre-trained language models for specialized tasks. These findings suggest promising avenues for developing more effective early intervention strategies in suicide prevention, while also highlighting the need for continued research to address limitations and ethical concerns in deploying such systems at scale.

Introduction

Suicide has only recently received widespread attention as a serious public health problem worldwide, accounting for an estimated 700,000 deaths annually [1].The American Psychological Association describes suicidality as: ”the risk of suicide, usually indicated by suicidal ideation or intent, especially as evident in the presence of a well-elaborated suicidal plan,” which can also present itself as the presence of suicidal ideas, plans, or attempts . Behavior or warning signs such as withdrawal, feelings of hopelessness, or self-harm are often part of the clinical assessment.

Over the past few years, increases in the uses of social media have changed communication and information exchange, offering access to the experiences of many people and real-time expressions of mental states. A survey in 2022 shows that 93% of U.S. teens use YouTube, 63% use TikTok, and 14% are users of Reddit [2]. As such, platforms such as Reddit, Twitter, and Facebook have become forums where individuals, especially young adults and teenagers, discuss their struggles with mental health more readily. However, several studies have established a correlation between social media use and mental health issues among teenagers. For instance, a study shows that while social media can provide support for teens, they also play a significant role in issues like cyberbullying, sleep disruption, as well as negative self-comparison from posts on those platforms, ultimately resulting in the development of mental health disorders such as depression and suicidal thoughts and behaviors [3,4]. Specifically, up to 57% of female high school students and 29% of male high school students experienced persistent feelings of sadness or hopelessness and 30% seriously considered attempting suicide [5].

Reddit is particularly valuable for researching suicide ideation detection due to its unique structural and cultural features. Unlike platforms where users are tied to real-world identities, Reddit allows for anonymity, enabling individuals to share personal and sensitive experiences more freely [6]. The platform's organization into topic-specific subreddits, such as r/SuicideWatch and r/teenagers, provides access to communities where discussions focused on one particular topic are prevalent [7]. Additionally, Reddit's upvoting and downvoting mechanisms influence the visibility of posts, reflecting community interest and engagement levels. This combination of anonymity, community focus, and user interaction makes Reddit a rich source of diverse expressions related to suicidal ideation.

Ethical considerations are paramount when using data from open forums like Reddit. This study adheres to ethical guidelines for using publicly available data [8]. Since Reddit posts are accessible to the public and users maintain anonymity through pseudonyms, the data collected did not involve identifiable personal information, and no private messages or deleted content were accessed. The potential for misinterpretation of user intent was recognized and therefore the analysis focuses on aggregate patterns rather than individual cases. The models developed are intended for research purposes and not for direct intervention

without professional oversight.

The vast amount of user-generated content on these platforms contains valuable insights into the mental health states of young individuals; hence, they offer the possibility of early detection and intervention in suicidal ideation. However, analyzing social media data for suicide risk assessment presents significant challenges due to the large volume of content generated daily. Platforms like Reddit host millions of posts and comments every day, not only making it impossible for human moderators to manually monitor all content but also raises ethical concerns regarding privacy and consent[9,10]. Furthermore, complexity arises from the diverse ways individuals express themselves online. The lack of standardization in language, including the use of slang, abbreviations, emojis, and cultural references, adds layers of complexity. People may express suicidal thoughts indirectly or through metaphors, sarcasm, or even humor, which can obscure their true intentions. This variability makes it difficult to develop models that accurately detect suicidal ideation across different contexts and user expressions.

Fortunately, recent advancements in deep learning and machine learning offer a more efficient and effective approach to analyzing large-scale textual data, enabling artificial intelligence models to effectively identify patterns and linguistic markers associated with suicidal thoughts, providing a promising tool for early detection and intervention[11]. Computational linguistics provides the tools to process and analyze natural language data, enabling the extraction of linguistic features such as syntax, semantics, and sentiment from text. These features serve as inputs for machine learning models, which learn patterns associated with suicidal ideation through training on labeled datasets.

Finally, the integration of AI in mental health monitoring, particularly in suicide prevention, becomes a very promising frontier of the field of digital health. Machine learning models from simple classifiers to sophisticated deep learning architectures have shown remarkable capabilities in natural language processing tasks. These models can be trained to recognize subtle linguistic cues, emotional undertones, and contextual nuances that may indicate an elevated risk of suicidal[12]. These AI models allow for the development of an automated system whereby, upon notification by the system, potentially alarming content is flagged for human review that may enable interventions in time to prevent fatal suicide attempts.

Previous research in this field has often relied on established models such as BERT (Bidirectional Encoder Representations from Transformers) for text classification tasks[13–15]. However, this study aims to explore novel deep-learning models that are less frequently tested in this context. Specifically, a hybrid Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) architecture[16] will be utilized alongside BERT to explore the effectiveness of these models on the evaluation of suicide ideation in Reddit posts. These models offer unique advantages in capturing both sequential dependencies and local features in text, potentially enhancing the accuracy of suicide risk detection.

This study is grounded in the theoretical framework of computational linguistics and machine learning applied to mental health detection. It builds upon the theory that linguistic patterns in social media posts can serve as indicators of mental health states, particularly suicidal ideation. This framework informs our methodology, specifically in guiding the selection of machine learning models and the interpretation of results in the context of suicide risk detection.

Results

Model Performance Comparison

The following models were compared for their potential in recognizing suicidal ideation: Decision Tree, LSTM-CNN, and BERT. Each model's performance was assessed using a range of metrics, with particular emphasis on the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), accuracy, precision, recall, F1-score, specificity, and sensitivity.

Among these three, the BERT model yielded the highest performance, achieving an AUC-ROC of 0.9909 and an accuracy of 97%. It demonstrated a high sensitivity of 96.76%, indicating a strong ability to correctly identify suicide-related posts, and a specificity of 95.93%, reflecting accurate detection of non-suicide-related content. The LSTM-CNN model achieved an AUC-ROC of 0.9611 and an accuracy of 90%, with a sensitivity of 88.59% and specificity of 90.60%. The Decision Tree baseline model showed reasonable performance with an AUC of 0.9477 and an accuracy of 88.16%.

The substantial performance increase achieved by the BERT model demonstrates the effectiveness of transformer-based architectures in capturing complex linguistic patterns. However, considering the high stakes of suicide risk detection, striving for even higher accuracy and robustness is essential if possible.

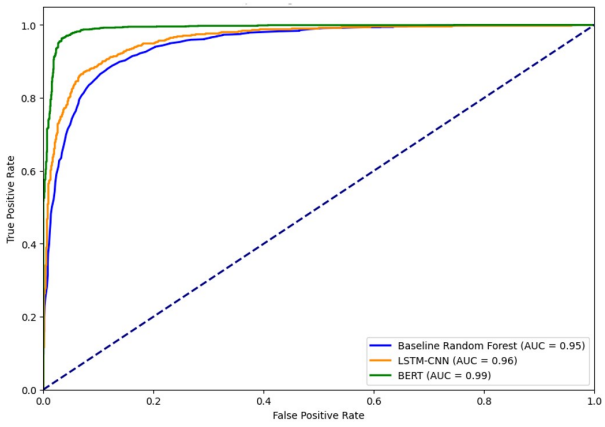

Fig. 1 Receiver Operating Characteristic (ROC) Curve

Model	Precision	Recall	F-1 Score	Accuracy	Specificity	Sensitivity
Decision Tree	0.88	0.88	0.88	0.88	0.8847	0.8784
LSTM-CNN	0.90	0.90	0.90	0.90	0.9060	0.8859
BERT	0.97	0.97	0.96	0.96	0.9593	0.9676

Table 1 Comparison of Performance Metrics for Models

Confusion Matrix Analysis

To provide a more granular understanding of model performance, confusion matrices were generated for each model. These matrices offer insights into the models' classification behavior across different post categories. For the LSTM-CNN model, out of 3,200 test samples, 1,421 were correctly identified as suicide-related posts (true positives), while 183 suicide-related posts were misclassified as non-suicide (false negatives). Additionally, 150 non-suicide posts were incorrectly flagged as suicide-related (false positives), and 1,446 were correctly identified as non-suicide (true negatives).

For the BERT model, out of 3,200 test samples, 1,552 were correctly identified as suicide-related posts (true positives), while 52 suicide-related posts were misclassified as non-suicide (false negatives). Additionally, 65 non-suicide posts were incorrectly flagged as suicide-related (false positives), and 1,531 were correctly identified as non-suicide (true negatives). The higher true positives and true negative rate compared to the LSTM-CNN model highlight the BERT model's superior sensitivity and specificity.

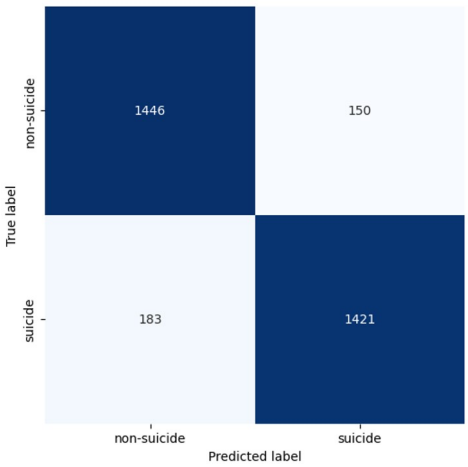

Fig. 2 Confusion Matric for LSTM-CNN Model

Feature Importance Analysis

Interpreting feature importance in deep learning models is challenging due to their complexity. To interpret the feature importance in our LSTM-CNN and BERT models, the Local Interpretable Model-agnostic Explanations (LIME) technique was employed[17]. LIME approximates the complex model locally with an interpretable model by perturbing the input data and observing changes in the output. This approach provided insights into the words contributing most significantly to the models' decision-making processes.

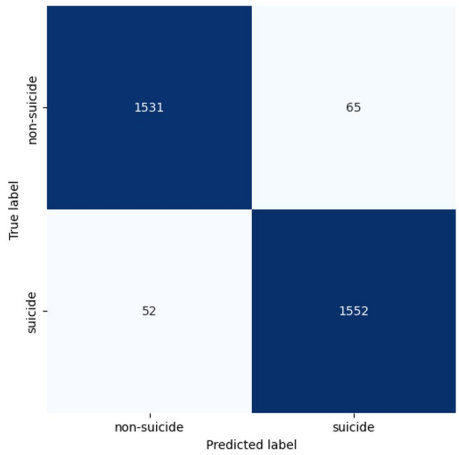

Fig. 3 Confusion Matrix for ERNIE Model

LIME revealed specific words that strongly impacted the models' assessments. For the suicidal text, terms like "suicidal," "suffer," "drugs," and "thoughts" had the highest positive contributions to the "suicidal" class. These words, associated with negative emotions, hopelessness, and references to self-harm or substance abuse, indicate that both models rely heavily on explicit expressions of distress and mental health struggles to classify a text as indicative of suicide risk.

Interestingly, the LIME visualizations showed that each model also captured subtle contextual cues. For example, phrases related to isolation and feelings of confinement (e.g., "confined") were highlighted as indicators of potential distress. The ability to detect these more nuanced expressions of emotional state demonstrates that both LSTM-CNN and BERT models are not merely counting specific keywords but are sensitive to context and the intensity of expressed emotions.

In comparison, non-suicidal texts were characterized by words with neutral or positive associations, such as "videogame" and "sup," which are expressions usually associated with positive emotions. In these cases, the absence of intense negative expressions and the presence of everyday conversational lan-

guage contributed to the models' classification of these texts as non-suicidal. However, it is important to note that the model classified "love" as contributing quite highly to the suicidal class. This suggests that some words with usually positive sentiment can also be associated with suicidal ideation when used in certain contexts. For example, the word "love" might appear in posts where individuals express feelings of unrequited love, loss of a loved one, or emotional pain related to relationships. In such cases, "love" becomes associated with expressions of despair. This highlights the models' sensitivity to contextual nuances, recognizing that words typically seen as positive can carry different emotional weights depending on how they are used.

Overall, LIME's interpretation offers valuable insight into how the models distinguish between suicidal and non-suicidal content. By identifying key linguistic markers of suicidal ideation, such as words that reflect hopelessness, substance abuse, and social isolation, LIME helps verify that the models are appropriately focused on relevant aspects of the text.

Sample suicidal text: intrusive suicidal thoughts. i'm 17, dependent on drugs, and i suffer from mdd. i con't know why i'm posting here, but i don't know where else to turn right now. it's constant and never stops.

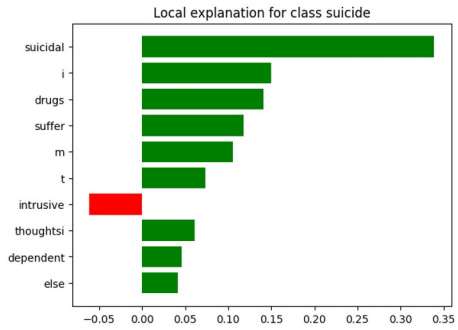

Fig. 4 LIME Explanation for LSTM-CNN

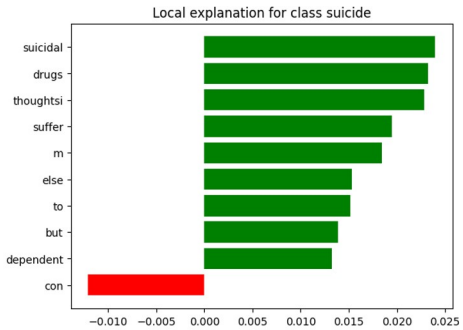

Fig. 5 LIME Explanation for BERT

Sample non-suicidal text: what's a great videogame i probably haven't heard of? sup, guys. i never post here, but i'm confined, not only to my house, but actually quarantined in my room, as i'm the only one in my house with covid. i'm so bored and i'd love something entertaining. i don't play a ton of videogames, so a new one would do pretty nicely. any suggestions? (pc only)

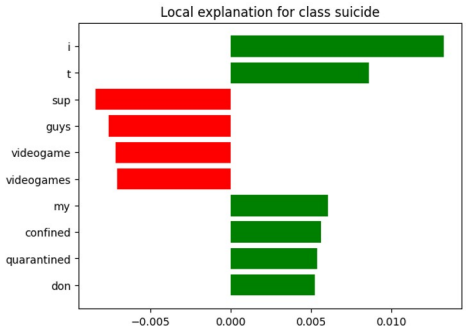

Fig. 6 LIME Explanation for LSTM-CNN

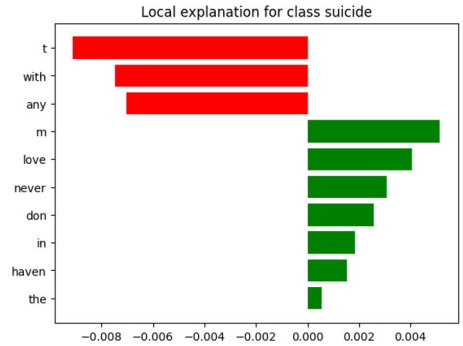

Fig. 7 LIME Explanation for BERT

Error Analysis

An error analysis of 100 sample misclassified texts revealed that the models sometimes misclassified posts expressing general distress as suicide-related and vice versa. For example, a post stating, "I'm so stressed about exams and feel like giving up," was incorrectly classified as suicide-related due to phrases like "giving up." Conversely, a post with subtle expressions like "I can't find a reason to keep going" was missed by the model, highlighting a lingering difficulty in capturing implicit cues. These examples illustrate the challenges in distinguishing between general negative sentiment and genuine suicidal ideation.

Hyperparameter Optimization Results

Optimal hyperparameters were identified for each model using Optuna[18]. For the LSTM-CNN model, the optimal configuration included a learning rate of approximately 0.0037, 179

LSTM units, 238 convolutional filters, a dropout rate of approximately 0.22, a batch size of 256, and 5 epochs. For the BERT model, the optimal configuration included a learning rate of approximately 1.53e-5, a batch size of 8, and 3 epochs. While more trials could potentially yield better hyperparameter configurations, the optimization was limited to 20 trials due to computational resource constraints and time considerations.

Discussion

The results of this study offer significant insights into the application of machine learning techniques for detecting suicide ideation in social media posts, specifically on Reddit. The BERT model's superior performance, achieving an AUC-ROC of 0.9909 and an accuracy of 97%, also demonstrated a high sensitivity of 96.76% and specificity of 95.93%, indicating a strong ability to correctly identify suicide-related or non-suicide-related posts. The BERT model's strong performance highlights the advantages of transformer-based architectures pre-trained on large English-language corpora. BERT's ability to capture contextual nuances and complex linguistic patterns makes it particularly effective for this task.

The LSTM-CNN model also performed well, achieving an AUC-ROC of 0.9611 and an accuracy of 90%. This suggests that hybrid architectures combining recurrent and convolutional layers can effectively capture both sequential dependencies and local features in text data. However, the performance gap between BERT and LSTM-CNN suggests that transformer models may offer a more powerful approach for detecting suicidal ideation.

The Decision Tree baseline model provided a reasonable starting point but was outperformed by the more complex models. With an AUC of 0.9477 and an accuracy of 88.16%, it demonstrated the limitations of simpler models in handling the complexity of natural language in the context of suicide ideation detection. This gap shows the importance of using advanced machine learning techniques to handle the complexity and subtlety inherent in suicide ideation detection.

One of the key findings of this study is the critical role of context in accurately identifying suicidal ideation. Traditional models like the Decision Tree rely heavily on keyword frequency and may miss the nuanced meanings behind certain phrases. The superior performance of BERT suggests that models capable of understanding context and semantic relationships can better interpret the nuanced language associated with mental health issues. For example, phrases like "I'm fine" can have different implications depending on the context and prior text, which BERT is better equipped to analyze.

The error analysis revealed common challenges across all models, particularly in dealing with ambiguous expressions and contextual nuances. For instance, posts expressing general distress or using colloquial language posed difficulties for the models. These difficulties highlight the inherent complexity of natural language and the subjective nature of interpreting mental health states from text alone. For example, sarcasm or idiomatic expressions like "I'm done with this" can be misinterpreted by models as indicators of suicidal ideation when they may simply express frustration. Conversely, subtle cries for help that don't use explicit language may be overlooked. Balancing sensitivity and specificity is crucial; while high sensitivity ensures that at-risk individuals are identified, elevated false positives can lead to unnecessary alarm and resource allocation. Implementing strategies such as adjusting classification thresholds, incorporating contextual awareness, and integrating external knowledge bases can help mitigate these issues. Additionally, incorporating human review for flagged content can ensure that nuanced cases are interpreted correctly.

The importance of minimizing false negatives also cannot be overstated in the context of suicide prevention. In our study, the BERT model had a sensitivity of 96.76%, meaning it correctly identified 96.76% of suicide-related posts. However, it still missed 3.24% of such posts. A missed indication of suicidal ideation could result in a lack of timely intervention, with potentially severe consequences. Therefore, models must prioritize sensitivity while maintaining acceptable specificity levels. Future models could incorporate ensemble methods or additional features, such as user engagement metrics, to enhance detection accuracy.

Limitations of this study include the potential biases introduced by the dataset. Moreover, the models' performance on a curated, balanced dataset may not directly translate to real-world scenarios where the prevalence of suicide-related posts is much lower. This raises important questions about the models' potential for false positives in practical applications. Future studies should evaluate these models on more diverse and imbalanced datasets to assess their robustness in realistic conditions. The dataset these models were trained on also solely used user posts from Reddit and may not be representative of other social media platforms such as Twitter, which often feature much shorter and informal text that may cause the models tested in these studies to underperform. Additionally, the non-suicide dataset being sourced from r/teenagers means it likely reflects only the language and concerns of adolescents, not adults who face a different set of struggles. This age-specific training data could limit the model's effectiveness across all age groups, potentially leading to reduced performance when applied to adult populations. Testing these models on data from diverse social media platforms and age groups would assess their robustness across different online contexts, communication styles, and life stages. More investigation should be done on data from other social media platforms and age demographics in order to find another model or build an ensemble model that performs well overall. Additionally, the models primarily analyze textual data; integrating multimodal data, such as images or user interaction

patterns, could provide a more comprehensive understanding of user behavior and mental state. Finally, the current study focused on static text classification and did not account for the temporal progression of suicidal ideation. Future research could explore sequential models that analyze a user's posts over time, potentially capturing the evolution of mental health states and providing earlier detection of at-risk individuals.

To address these limitations and advance the field, future research should consider several key directions. Expanding the training data to include a wider range of social media platforms and demographic groups could improve model generalization and reduce potential biases. Incorporating other forms of user data, such as posting frequency, interaction patterns, and profile information, could provide a more holistic view of an individual's mental state. Developing models that can track changes in linguistic patterns over time might enable earlier intervention by identifying gradual shifts toward suicidal ideation.

Implementing advanced methods for model interpretation could bridge the gap between performance, explainability, and ethical considerations. However, it is important to note that ensuring user privacy and data security is essential, necessitating adherence to regulations like GDPR and respect for user consent. Transparent policies about data use, options to opt-out, and safeguards against misuse are critical to maintain trust and protect individuals' rights. Moreover, there is a need to address the potential psychological impact on users whose content is flagged, as misclassification could lead to unwarranted interventions or stigmatization. Developing guidelines and involving mental health professionals in the review process can help mitigate these risks.

Further research could explore the integration of these models into mental health support systems, potentially facilitating early detection and timely interventions. However, integrating these machine learning models into existing mental health support systems presents both opportunities and challenges. Automated detection can facilitate early identification of at-risk individuals, potentially enabling timely support and intervention. However, reliance on automated systems must be balanced with human oversight to interpret nuanced cases appropriately. Collaboration with mental health practitioners is essential to ensure that the models are used as supportive tools rather than definitive diagnostic systems.

In conclusion, this study demonstrates the potential of machine learning, particularly hybrid neural network architectures like LSTM-CNN, in detecting suicide ideation from social media posts. The findings contribute to the growing body of research on AI applications in mental health, offering both promising avenues and cautionary insights. As research in this field progresses, maintaining a balance between model performance, interpretability, and ethical considerations will be very important. The goal remains to develop tools that can effectively support mental health professionals and individuals at risk while respecting the deeply personal nature of mental health challenges. Future work in this domain has the potential to significantly impact early intervention strategies and ultimately contribute to suicide prevention efforts on a broader scale.

Methods

Dataset Acquisition and Characteristics

This study drew from a rich dataset of Reddit posts, specifically targeting content from the r/SuicideWatch subreddit. The Pushshift API was selected for data collection due to its extensive archive of Reddit posts and comments, enabling researchers to retrieve large datasets efficiently [19]. It provides historical data that may not be accessible through Reddit's official API, which is crucial for longitudinal studies. However, Pushshift may have limitations, such as occasional delays in data indexing and potential exclusion of posts removed by moderators before archiving. To create a balanced dataset, posts from r/SuicideWatch were labeled as suicide-related, while non-suicide posts were sourced from the r/teenagers subreddit. Posts from r/teenagers were used for non-suicide-related posts to align with the focus on adolescent populations, ensuring that the language style and topics are age-appropriate for comparison. However, it is acknowledged that this may introduce bias, as the subreddit might not reflect the broader Reddit user base and therefore the linguistic patterns and concerns expressed may differ from those of adults or other demographic groups. This dataset accumulated a total of 348,110 posts of relatively equal distribution between suicide-related and non-suicide-related content [20].

Data Preprocessing and Feature Engineering

The raw text data underwent a preprocessing process before being utilized. After removing null values and standardizing all text to string format, the potential class imbalance was addressed by randomly sampling 8,000 posts from each category using stratified sampling, accumulating to 16,000 posts total. This step ensured a manageable yet representative dataset for model training, as the full dataset of over 300,000 posts was deemed excessive for this task and would lengthen training times without significantly improving accuracy. The resulting dataset comprised 16,000 posts, with 8,000 labeled as suicide-related and 8,000 as non-suicide-related, achieving a 1:1 ratio.

Text preprocessing included lowercasing, removal of URLs, mentions, and special characters. Sentiment analysis was performed using the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool to assess the emotional tone of posts [21]. Our analysis focuses on several linguistic patterns known to correlate with mental health states. Sentiment analysis was performed to assess the emotional tone of posts, identify the use

of first-person singular pronouns, and detect negative emotion words, which are associated with depression and suicidality[22].

Model Architectures and Implementation

1. **Baseline Decision Tree Classifier:**
   A Decision Tree classifier, implemented using scikit-learn[23], served as the baseline model. This approach vectorized the text data using Term Frequency-Inverse Document Frequency (TF-IDF), limiting the feature set to the 5,000 most significant terms. TF-IDF was chosen for vectorizing text data due to its ability to weigh terms based on their importance within a document relative to the entire corpus. This method reduces the impact of commonly used words that may not be informative for classification purposes. Limiting the feature set to the top 5,000 terms based on TF-IDF scores helps reduce dimensionality and computational load while retaining the most informative features. The decision tree was configured with a maximum depth of 20 to prevent overfitting while still allowing for complex decision boundaries. To balance the model's complexity and generalization ability, the maximum depth was set based on cross-validation results, which showed that deeper trees began to overfit the training data without improving validation performance.

2. **LSTM-CNN Hybrid Model:**
   The LSTM-CNN model represented a more sophisticated approach that was implemented using TensorFlow and Keras[24], leveraging the strengths of both recurrent and convolutional neural networks. This model began with an embedding layer initialized with pre-trained GloVe embeddings (100-dimensional)[25]. To preserve the valuable information encoded in these embeddings, this layer remained frozen during training.

   The architecture's core consisted of parallel LSTM and CNN branches. The LSTM component excelled at capturing long-range dependencies in the text, while the CNN extracted local features. These complementary outputs were then merged through a concatenation layer. The model finished with dense layers using ReLU activation, dropout layers for regularization, and a final dense layer with softmax activation for binary classification.

   Hyperparameters such as the number of LSTM units, convolutional filters, dropout rate, learning rate, batch size, and number of epochs were optimized using Optuna's Bayesian optimization in a total of 20 trials. Early stopping was implemented based on validation loss to halt training when no improvement was observed for two consecutive epochs, preventing overfitting.

3. **ERNIE-based Model:**
   The BERT model was implemented to test the effectiveness of pre-trained transformer models on the task. The 'bert-base-uncased' model from Hugging Face Transformers was used[26]. BERT has been pre-trained on large English-language corpora, making it suitable for the dataset used in this study. Hyperparameters such as learning rate, batch size, and number of epochs were optimized using Optuna, with 20 trials conducted. Early stopping was also implemented, similar to the LSTM-CNN model, to prevent overfitting. The model was fine-tuned on the dataset with appropriate adjustments to capture domain-specific nuances.

Hyperparameter Optimization

Bayesian optimization via Optuna was applied for both the LSTM-CNN and ERNIE models. This approach allowed for a systematic exploration of the hyperparameter space, with 20 trials conducted for each model. The ERNIE optimization ran for 1 hour, while the LSTM-CNN optimization ran for 13 minutes. The optimization process focused on maximizing validation accuracy and explored a range of hyperparameters. Hyperparameter optimization sampled the learning rate on a logarithmic scale between 1e-5 and 1e-2, chose batch sizes from 32, 64, or 128, and set the number of training epochs between 5 and 8. Model-specific parameter tuning varied the number of LSTM units and CNN filters from 32 to 256 and adjusted dropout rates to between 0.1 and 0.5.

Training Protocol and Evaluation Metrics

The dataset was partitioned into training (80%) and testing (20%) sets using stratified sampling to maintain the class distribution. For the deep learning models, the Adam optimizer was employed with a dynamic learning rate schedule. This schedule reduced the learning rate by a factor of 0.1 when the validation loss plateaued, allowing for fine-grained optimization as training progressed.

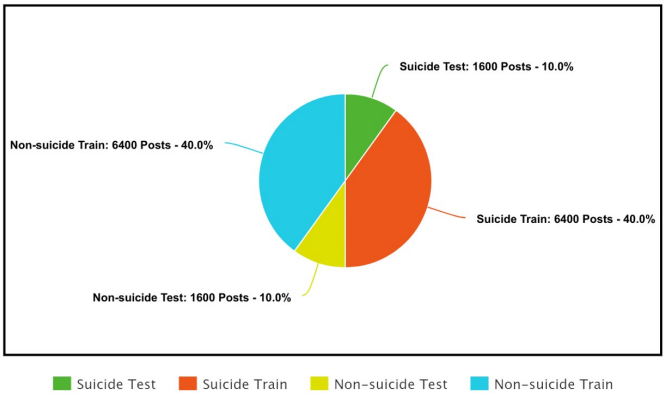

Fig. 8 Data Distribution

An early stopping mechanism was implemented with a patience of 5 epochs, monitoring validation loss to guard against overfitting. Model checkpointing saved the best-performing model based on validation accuracy, which ensured the final model represented the peak of performance rather than potentially overfitted iterations.

The models were evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, specificity, sensitivity, and area under the Receiver Operating Characteristic curve (AUC-ROC). Sensitivity (recall) is particularly important in suicide ideation detection, as it measures the model's ability to correctly identify true positives. High sensitivity reduces the risk of false negatives, which is crucial since missing an at-risk individual could have severe consequences. Specificity assesses the model's ability to correctly identify true negatives, helping to minimize false positives and avoid unnecessary interventions. Finally, confusion matrices were generated to provide a detailed visual view of model performance across classes.

Computational Environment

All experiments were conducted in a cloud environment using Google Colab, utilizing its T4 GPU for accelerated computing. The deep learning models were implemented and trained using TensorFlow 2.5.0 and PyTorch 1.9.0, while scikit-learn 0.24.2 was employed for the baseline Decision Tree model and for calculating various evaluation metrics.

Acknowledgments

I would like to express my sincere gratitude to Xi Zhou for her valueable and helpful suggestions on this project. Her advice greatly expanded my knowledge and helped me through the difficult beginning stages of this project by pointing out possible approaches.

References

1 *World Health Organization. Suicide*, https://www.who.int/news-room/fact-sheets/detail/suicide.

2 S. Atske, *Teens and Social Media Fact Sheet*, https://www.pewresearch.org/internet/fact-sheet/teens-and-social-media-fact-sheet/.

3 M. Choudhury and E. Kiciman, *The language of social support in social media and its effect on suicidal ideation risk*.

4 *American Psychological Association*, https://psycnet.apa.org/buy/2022-19150-001.

5 *Centers for Disease Control and Prevention*, https://www.cdc.gov/media/releases/2023/p0213-yrbs.html.

6 P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel and M. Strohmaier, *Evolution of Reddit: From the front page of the internet to a self-referential community?*

7 Reddit, *Reddit*, https://www.reddit.com/.

8 L. Townsend and C. Wallace, *Social Media Research: A Guide to Ethics*.

9 S. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*.

10 P. Services, *Reconciling Statistical and Clinicians' Predictions of Suicide Risk*, https://psychiatryonline.org/doi/full/10.1176/appi.ps.202000214.

11 G. Coppersmith, R. Leary, P. Crutchley and A. Fine, *Natural language processing of social media as screening for suicide risk*.

12 A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement and Z. Kaminsky, *A machine learning approach predicts future risk to suicidal ideation from social media data*.

13 P. Boonyarat, D. Liew and Y.-C. Chang, *Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19*.

14 J. Gorai and D. Shaw, *A BERT-encoded ensembled CNN model for suicide risk identification in social media posts*, https://doi.org/10.1007/s00521-024-09642-w.

15 S. Devika, M. Pooja, M. Arpitha and R. Vinayakumar, *BERT-based approach for suicide and depression identification*.

16 J. Zhang, Y. Li, J. Tian and T. Li, *LSTM-CNN hybrid model for text classification*.

17 M. Ribeiro, S. Singh and C. Guestrin, *Why should I trust you?": Explaining the predictions of any classifier*.

18 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*.

19 J. Baumgartner, S. Zannettou, B. Keegan, M. Squire and J. Blackburn, *The Pushshift Reddit dataset*.

20 N. Komati, *Suicide and Depression Detection*, https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch.

21 C. Hutto and E. Gilbert, *VADER: A parsimonious rule-based model for sentiment analysis of social media text*.

22 J. Pennebaker, M. Mehl and K. Niederhoffer, *Psychological aspects of natural language use: Our words, our selves*.

23 scikit learn, *scikit-learn: machine learning in Python — scikit-learn 1.5.1 documentation*, https://scikit-learn.org/stable/.

24 M. Abadi, A. Agarwal and P. Barham, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*, https://www.tensorflow.org/static/extras/tensorflow-whitepaper2015.pdf.

25 J. Pennington, *GloVe: Global Vectors for Word Representation*, https://nlp.stanford.edu/projects/glove/.

26 T. Wolf, L. Debut and V. Sanh, *Transformers: State-of-the-art natural language processing*.

