Detection of Suicidal Risk on Social Media: A Hybrid Model

Zaihan Yang, Ryan Leonard, Hien Tran, Rory Driscoll, and Chadbourne Davis

Department of Math and Computer Science, Suffolk University, Boston, USA
zyang13@suffolk.edu
{rleonard2, hien.tran, rdriscoll, chad.davis}@su.suffolk.edu

Abstract. Suicidal thoughts and behaviors are an urgent public health concern, underscoring the need for effective tools to enable early detection of suicide risk. We address this challenge by developing robust machine learning models that classify Reddit posts into four distinct suicide risk severity levels. Framing this as a multi-class classification task, we propose a RoBERTa–TF-IDF–PCA Hybrid model that integrates deep contextual embeddings from Robustly Optimized BERT Approach (RoBERTa) with statistical term-weighting from TF-IDF, whose features are reduced via Principal Component Analysis (PCA) to enhance accuracy and stability. To mitigate data imbalance and overfitting, we explore a range of data resampling and data augmentation strategies to improve model generalization. We compare our hybrid approach against RoBERTa-only, BERT, and traditional machine learning classifiers. Experimental results demonstrate that our model can achieve improved performance, giving a best weighted F_1 score of 0.7512.

Keywords: Classification, Deep Learning, Machine learning, Transformers, Large Language models, data imbalance, data augmentation, Evaluation, Mental Health, Suicidal Ideation Detection

1 Introduction

Suicidal thoughts and behaviors are a growing societal concern. According to the World Health Organization, approximately 700,000–800,000 people die by suicide globally each year. In the U.S., suicide is the second leading cause of death among individuals aged 10–34 and the fourth among those aged 35–64. Suicidal ideation varies in intensity—from persistent thoughts, to active planning or engagement in self-harm behaviors (e.g., cutting, burning), to actual suicide attempts (e.g., overdose, jumping, firearms). Multiple factors contribute to mental health decline and suicidal outcomes[1], many of which can be mitigated through timely interventions.

Both psychologists and computer scientists have developed tools for early detection of suicide risk. Psychologists typically rely on self-reported questionnaires, interview transcripts, and clinical records, using statistical analysis to identify risk factors[2, 3], whereas computer scientists increasingly apply machine or deep learning techniques to social media data. Platforms such as Twitter (X)[13], Reddit[7–15, 19], Tumblr[17], and ReachOut[26] provide abundant textual content, with Reddit in particular offering longer, more structured, and context-rich discussions. Its anonymity and openness encourage candid sharing of personal struggles, but also introduce challenges: privacy concerns, scarcity of labeled data, linguistic diversity across cultures, and noisy or misleading posts. Moreover, identifying posts reflecting different severity levels remains difficult, as severe ideation is less frequently expressed publicly, leading to class imbalance.

Research on suicidal risk detection has explored a wide range of modeling strategies. Early work relied on traditional machine learning classifiers such as SVM, Logistic Regression, Random Forest, and XGBoost[10, 19], paired with feature engineering techniques like bag-of-words, TF-IDF scores, n-grams, lexicons, and emotion indicators[1]; metadata such as age or gender has also been used to improve classification[5, 18]. More recent approaches employ deep learning architectures—CNN, RNN, LSTM[7, 8, 11–15]—and, increasingly, transformer-based language models such as BERT, RoBERTa, Gemma, GPT, and LLaMA[9, 14, 16, 20–22], which excel at capturing deep contextual meaning. While transformers generally outperform traditional models in understanding complex and nuanced linguistic patterns, they require substantial computational resources and often lack interpretability. Hybrid methods that combine deep, transformer-based contextual embeddings with engineered features remain underexplored but hold promise for enhancing both performance and interpretability.

Risk severity modeling ranges from binary classification (suicidal vs. non-suicidal)[10, 13, 15] to multi-class frameworks[8, 9, 11, 22]. Notable examples include four-level schemes (no risk, low, moderate, severe)[11] and variations (indicator, ideation, behavior, attempt)[8, 9]. These differing formulations highlight the ongoing need for robust, fine-grained severity detection.

To address the above challenges, we propose a hybrid model that integrates RoBERTa-based word embeddings with TF-IDF features. RoBERTa captures semantic context, while TF-IDF emphasizes statistically salient terms. Our contributions include:

- Proposed a RoBERTa–TF-IDF hybrid model that integrates RoBERTa-based contextual embeddings with TF-IDF's statistical weighting;
- Applied dimensionality reduction to TF-IDF vectors using Principal Component Analysis (PCA) to improve efficiency and reduce noise;
- Collected and manually annotated Reddit posts into four suicidal risk severity levels;
- Explored various data resampling and augmentation techniques to address class imbalance and overfitting;
- Conducted a comparative analysis of our hybrid model against RoBERTa, BERT, and traditional classifiers, demonstrating superior performance.

2 Data Collection and Annotation

2.1 Data Set

Our dataset combines two primary sources. First, we used a dataset from researchers at The Hong Kong Polytechnic Universitywhich includes 500 labeled and 1,600 unlabeled Reddit posts from 14 mental health-related subreddits (e.g., r/SuicideWatch, r/depression). These posts are written in natural English and contain typical online artifacts like typos, emojis, and occasional non-English characters.

Second, we conducted our own data scraping using Python's Reddit API wrapper (PRAW), targeting r/SuicideWatch. Posts were collected using various sorting methods (e.g., "top", "new", "hot") across multiple time frames to ensure diversity and reduce duplication. To retain context, we filtered for posts where the original author had commented. This yielded 899 posts from 473 users, dated between 2008-12-16 and 2025-01-03. In total, our dataset consists of 2,999 posts. Table 1 presents summary statistics. Note that word token counts include English words, numbers, words with underscores, and Latin-based non-English words, while excluding symbols and punctuation.

Table 1: Data Set Statistics

Number of posts	**Number of Users**	**Number of Distinct Word Tokens**	**Average Number of tokens per post**
2999	473	13062	150.27

2.2 Ground-truth human-labeling

We adopted the labeling framework and annotation criteria validated by psychology experts, as outlined in[8], and also used by The Hong Kong Polytechnic University for their 500-post dataset[9]. The risk level definitions are summarized in Table 2.

Using these guidelines, we annotated the remaining 2,499 posts. The annotation team included one faculty member and four undergraduates, all in Computer Science. Inter-annotator agreement, measured by Fleiss's Kappa, was 0.5641, indicating moderate reliability.

Figure 1 shows the distribution of posts across risk levels. The data is imbalanced, with "ideation" accounting for 45.28%, followed by "indicator" (28.21%), "behavior" (17.97%), and "attempt" (8.54%). This imbalance complicates classification, especially for underrepresented categories. To address this, we experimented with re-sampling and data augmentation techniques.

Table 3 provides example posts for each risk level, highlighting distinct linguistic and conceptual traits across categories.

Table 2: Definition of Suicidal Risk Levels

Category(Risk-Level)	**Definition**
Indicator	The post content has no explicit expression concerning suicide
Ideation	The post content has explicit suicidal expression but there is no plan to commit suicide
Behavior	The post content has explicit suicidal expression and a plan to commit suicide or self-harming behaviors
Attempt	The post content has explicit expressions concerning historic suicide

Fig. 1: Distribution of Suicidal Risk Levels

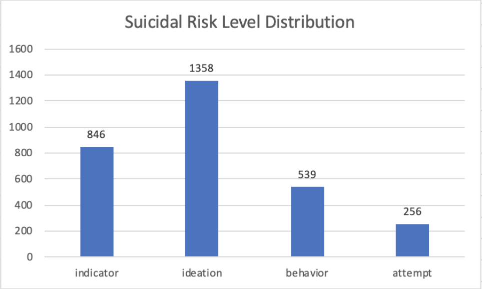

Table 3: Four suicidal risk level Examples

Indicator: *People who commit suicide are not weak. Most of the time people get to that place because we've bottled everything up and tried to deal with painful emotions by ourselves then one day we just snap and break down.*

Ideation: *Where can I find out how to die. I don't care if it's slow or painful or both, I just want to definitely die*

Behavior: *"I'm going to kill myself in 6 months if things doesn't get better. I can't be feeling this way forever and I'm so tired of struggling with my mental health. If things doesn't get better then what's the point of being alive."*

Attempt: *I hope I don't make it. I don't want to survive another attempt because all my parents will tell me is how expensive the medical bill is. I just want to die without having to hear them again.*

3 Model Design

Our proposed hybrid model integrates the word-embedding learned from the Robustly Optimized BERT Approach (RoBERTa)[24], one of the state-of-the-art transformer models extending BERT[23], with the statistical term-weighting of TF-IDF (Term Frequency-Inverse Document Frequency). We also explored different data resampling as well as data augmentation techniques to deal with the problem of imbalanced data and overfitting.

3.1 the RoBERTa-TF-IDF-PCA Hybrid Model

The working mechanism of our hybrid model is detailed below and illustrated in Figure 2.

- ***Tokenization***: RoBERTa uses byte-level Byte-Pair Encoding (BPE) to tokenize each post, converting tokens into input IDs and attention masks. Input IDs are unique integers from the pre-trained model's vocabulary, while attention masks are binary (1 for real tokens, 0 for padding) to indicate which tokens the model should attend to. A single *[CLS]* token is added at the beginning of the sequence for the entire post. For long posts, tokens beyond the 512-token limit (for BERT-base and RoBERTa-base) are truncated.
- ***Embedding***: After tokenization, input IDs are passed through an embedding layer, which maps each token ID to a dense and fixed-length vector representation (embedding) with positional embedding added to provide information about the position of each token in the sequence. In both the RoBERTa and BERT base models, the fixed length is set to be 768.
- ***Transformer Layers***: This embedding (512 * 768 tensor) is then passed through 12 Transformer layers, where it undergoes iterative updates via self-attention and feedforward layers, refining the representation in the context of the entire sequence. The embedding of the *[CLS]* token is finally generated and extracted representing the aggregated information for the entire post.
- ***TF-IDF and PCA Integration***: In addition to contextual embeddings from RoBERTa, we incorporate feature vectors derived from TF-IDF representations of the original text. We first represent each tokenized post as a vector of the M most significant features based on their TF-IDF scores. To reduce noise and vocabulary dimensionality, we further applied Principal Component Analysis (PCA) to the TF-IDF vectors, selecting the top N components that capture the most variance. The resulting N-dimensional TF-IDF-PCA vector is then concatenated with the *[CLS]* token embedding obtained from the final transformer layer. This yields a combined feature representation of dimensionality $768 + N$, where 768 corresponds to the RoBERTa *[CLS]* embedding. In our experiments, we set M be the number of unique tokens in the dataset and N be 300 based on empirical validation.
- ***Output Layer and Loss Function***: The enriched post-level embedding of the *[CLS]* token with the size of $768 + N$ will then be passed through a linear layer followed by a *softmax* activation to predict the probability distribution of different categories of classes (suicide risk levels) $P(Y_R^i) = softmax(linear(\hat{x^i}))$, where the *softmax* function normalizes the output to a [0, 1] range. With $P(Y_R^i)$ computed, categorial cross entropy loss is then adopted for updating of the model parameters, i.e., $L_R = \frac{1}{D}\sum_{i=1}^{D}\hat{Y_R^i} \cdot \log(P(Y_R^i))$, where D is the total number of posts in the training set and $\hat{Y_R^i}$ is the ground-truth label of post i. The cross-entropy loss

is minimized using the $AdamW$ optimizer, which updates the model's parameters through gradient descent.

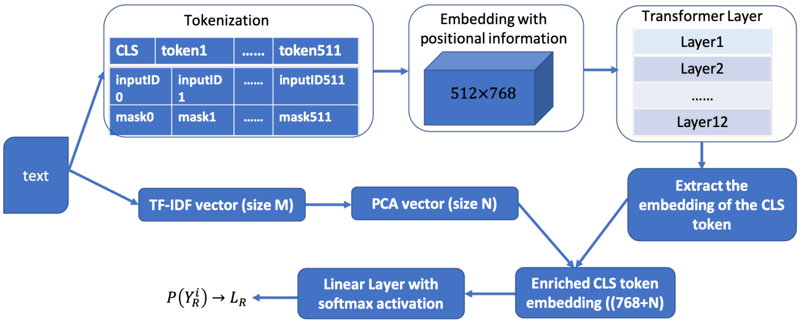

Fig. 2: Hybrid Model Framework

3.2 Data Resampling

As indicated in Figure 1, the data is imbalanced across the four suicidal risk categories. The "ideation" class, reflecting general suicidal thoughts without plans or actions, contains far more posts than the "attempt" class, which involves actual suicide attempts. However, identifying "attempt" cases is especially critical for timely intervention. To explore whether balanced training data improves classification, we applied three resampling techniques: 1) ***Over-Sampling (OSam)***: Based on the class with the most posts, we duplicate instances from other classes to balance the training set; 2) ***Under-Sampling (USam)***: Based on the class with the fewest posts, we randomly reduce instances from other classes; 3) ***Sample Weighted Loss Function (SWL)***:We apply a weighted loss using the inverse class proportions: $L_R = \frac{1}{D}\sum_{i=1}^{D}\frac{D}{D_{\hat{Y}_R^i}}\hat{Y}_R^i \cdot \log(P(Y_R^i))$, where $D_{\hat{Y}_R^i}$ is the number of instances in category i.

3.3 Data Augmentation

Data augmentation in NLP involves altering text while preserving its meaning to boost model robustness. Some techniques also expand the dataset by generating lexically similar samples, improving generalization. We explored the following methods: 1) ***Extending abbreviated words***: Reddit posts often use abbreviations (e.g., "kms" for "kill myself", "idk" for "I do not know"). To standardize language and improve clarity, we created a dictionary of 111 abbreviation-expansion pairs and applied it to training data. 2) ***Extending emojis into text***: Among the 2,999 posts, 46 contain emojis that convey emotion. We replaced them with corresponding English descriptions to enhance contextual understanding. 3) ***Text summarization***: 142 posts exceed 512 tokens—the input limit for RoBERTa-base and BERT-base—leading to truncation and potential loss of meaning. We applied summarization to retain core content within this limit. 4) ***Using Google Translate***: Posts were translated to Spanish and back to English, introducing variation in phrasing while preserving meaning. This increased data diversity and helped improve generalization.

4 Experimental Results

4.1 Experiment Setups

We used five-fold cross-validation to ensure stable results. The 2,999 posts were split into five non-stratified folds, allowing each to reflect varied class distributions, which better simulates real-world conditions. In each iteration, four folds were used for training and validation, and one for testing. The training-validation set was further split 80:20 (stratified) to create separate subsets. Models were trained with early stopping to prevent overfitting on the validation set, and then evaluated on the test fold.

Classification performance was measured by comparing predicted labels with ground truth, and results were averaged over all five folds for generalization. Given the class imbalance (Figure 1), we used weighted precision, recall, and F_1 scores for evaluation.

4.2 Experiment Results

We made several groups of comparisons in terms of classification results between different models, data resampling techniques and data augmentations.

Data Resampling During five-fold cross-validation, we applied oversampling, undersampling, or a weighted loss function to balance the training and validation sets, ensuring equal post counts across risk levels. The test set distribution remained unchanged in all experiments. Table 4 shows the results, where "original" denotes the use of unaltered data distributions without resampling.

Table 4: Comparision of Different Data Resampling Techniques

	weighted precision	**weighted recall**	**weighted F_1 score**
Original	0.7523	**0.7496**	**0.7499**
OSam	0.7485	0.7439	0.7421
USam	0.7120	0.7009	0.7022
SWL	**0.7524**	0.7486	0.7480

As shown, the Original, Oversampling (OSam), and Sample Weighted Loss (SWL) methods performed similarly, with the Original method achieving the highest weighted F_1 score of 0.7499. This may be due to the non-stratified fold split, which introduced variability in class distribution across folds.

OSam balanced training data but may have introduced a mismatch with the imbalanced test distribution, affecting generalization. Similarly, SWL assigned inverse-proportional weights, which may not align with the test set's distribution. Undersampling yielded the lowest performance, likely due to reduced training data, limiting the model's learning capacity.

Data Augmentation In this group of comparison, we incorporated all four data augmentation techniques (with DA) we introduced in Section 3.3 into RoBERTa learning process, and compared the classification results against the model without data augmentation (w/o DA). The original data distribution was used in this experiment, and data augmentation was applied by modifying the text of the posts without increasing the dataset size. Table 5 shows the results.

Table 5: With vs. Without Data Augmentation

	weighted precision	**weighted recall**	**weighted F_1 score**
w/o DA	0.7523	0.7496	0.7499
DA	0.7385	0.7342	0.7341

As indicated, incorporating data augmentation techniques did not improve classification performance, as the weighted precision, recall, and F_1 scores all decreased. Several factors may contribute to this decline: Expanding abbreviations and emojis can lead to unnatural sentence structures that are misaligned with RoBERTa's learned embeddings. Since abbreviations and emojis often carry contextual or sentiment cues, expanding them may dilute or distort their meaning; Text summarization may slightly alter the original intent of a post, potentially affecting classification accuracy; Google Translate can introduce inaccuracies or unnatural synonyms, making sentences less representative of the original text. Further investigation is needed to analyze the impact of each augmentation technique individually.

Hybrid Model We compared our hybrid model against RoBERTa-only and BERT-based baselines. Results are shown in Table 6.

Table 6: Performance of RoBERTa-TF-IDF-PCA: the Hybrid Model

	weighted precision	**weighted recall**	**weighted F_1 score**
RoBERTa-TF-IDF-PCA	**0.7557**	**0.7532**	**0.7512**
RoBERTa	0.7523	0.7496	0.7499
BERT	0.7045	0.6996	0.6949

For both the hybrid RoBERTa-TF-IDF-PCA model and standalone RoBERTa, we used the original class distribution without resampling or data augmentation, as these yielded the best results (see Table 4). BERT(DistilBERT) was included as an additional baseline. Hyperparameters were set empirically: batch size = 3, learning rate = 1e-5, and weight decay = 0.01 (L2 regularization). Notably, the TF-IDF features were computed globally using unigrams extracted from the entire dataset, with no minimum or maximum document frequency thresholds applied for filtering. The number of PCA components is set to be 300 empirically.

The hybrid model outperformed all others, achieving the highest weighted scores in precision, recall, and F_1, with a top F_1 score of 0.7512.

Table 7: Classification results for four severity levels

Model	**Indicator-F_1**	**Ideation-F_1**	**Behavior-F_1**	**Attempt-F_1**
RoBERTa-TF-IDF-PCA	0.7541	0.7961	0.6767	0.6527
RoBERTa	0.7681	0.7875	0.6721	0.6437

Table 7 shows the average F_1 score per risk level. The hybrid model outperforms RoBERTa alone in three of four categories. Performance aligns with class frequency:

ideation scores highest, followed by *indicator*, *behavior*, and *attempt*, reflecting the effects of class imbalance and the need for targeted modeling strategies.

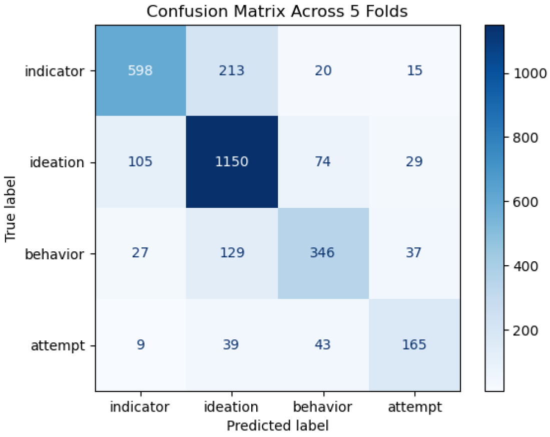

Fig. 3: Confusion Matrix for four different suicidal risk severity

Figure 3 shows the confusion matrix across five folds. Frequent misclassifications occur between *indicator* and *ideation*, and *behavior* is often confused with both *ideation* and *attempt*. These patterns suggest the model struggles to separate conceptually similar categories. Improved results may require more training samples and better discriminative features.

Comparison with Traditional Machine Learning Models Before the emergence of deep learning models like BERT and RoBERTa, traditional machine learning classifiers performed well on many classification tasks. In this study, we evaluated their effectiveness for suicidal risk detection and compared them with RoBERTa. We tested four classifiers: Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and Random Forest, using two feature types: 1) ***TF-IDF***: Posts were tokenized and represented using the 3,000 most significant TF-IDF features. 2) ***Word2Vec***: Posts were tokenized and encoded into 1,000-dimensional vectors using Word2Vec[25], where semantically similar words are close in vector space.

With TF-IDF, we also evaluated different pre-processing methods (stopword removal, stemming, lemmatization) to assess their impact. Results in Table 8 were obtained via five-fold cross-validation and evaluated using the weighted F_1 score.

Table 8: Comparison with traditional classifiers

	SVM	**Logistic Regression**	**Naive Bayes**	**Random Forest**
TF-IDF	0.5849	0.5658	0.3073	0.4717
TF-IDF + remove stopwords	0.5727	0.5637	0.3227	0.5033
TF-IDF + stemming	0.5880	0.5725	0.3061	0.4805
TF-IDF + lemmatization	0.5825	0.5725	0.3082	0.4711
Word2Vec	0.4347	0.4901	0.3387	0.4831

Key findings include: 1) SVM consistently outperformed other traditional models, followed by Logistic Regression, Random Forest, and Naive Bayes (using Gaussian with Word2Vec, Multinomial with TF-IDF). 2) TF-IDF outperformed Word2Vec across all classifiers. 3) Pre-processing effects varied: stemming helped SVM and Logistic Regression; stopword removal improved Naive Bayes and Random Forest. 4) As shown in Table 6, RoBERTa surpassed all traditional classifiers, underscoring the strength of transformer-based models for this task.

5 Summary and Future work

Timely detection of varying levels of suicidal ideation is essential for effective intervention and suicide prevention. We frame this task as a multi-class classification problem and propose a hybrid RoBERTa–TF-IDF–PCA model that combines RoBERTa embeddings with TF-IDF vectors. To mitigate data imbalance and overfitting, we employed various resampling and augmentation techniques. We also compared traditional classifiers using different feature sets against RoBERTa-based models. Our hybrid approach achieved the best performance, with a top weighted F_1 score of 0.7512.

Future directions include: (1) exploring additional ensemble methods to enhance generalization; (2) testing alternative combinations of feature representations; (3) evaluating other deep learning models, including LSTM and newer transformers like Gemma and LlaMA; (4) incorporating richer context such as follow-up comments or related posts; and (5) conducting temporal analysis of user posts to better track changes in risk over time for personalized interventions.

References

1. Ji S, Pan S, Li X, Cambria E, Long G and Huang Z (2021). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. IEEE Transactions on Computational Social Systems, 8(1), pp. 214-226.
2. Ati NAL, Paraswati MD and Windarwati HD (2021). What are the risk factors and protective factors of suicidal behavior in adolescents? A systematic review. Journal of Child and Adolescent Psychiatric Nursing, 34(1), pp. 7-18.
3. Castillo-Sánchez G, Marques G, and Dorronzoro E (2020). Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. Journal of Medical Systems, 44(12), 205.
4. Heckler, W. F., de Carvalho, J. V., and Barbosa, J. L. V. (2022). Machine learning for suicidal ideation identification: A systematic literature review. Computers in Human Behavior, 128, Article 107095.
5. Ehtemam, H., Sadeghi Esfahlani, S., Sanaei, A. et al. Role of machine learning algorithms in suicide risk prediction: a systematic review-meta analysis of clinical studies. BMC Med Inform Decis Mak 24, 138 (2024).
6. Raymond Su, James Rufus John, Ping-I Lin, Machine learning-based prediction for self-harm and suicide attempts in adolescents,Psychiatry Research,Volume 328,2023,115446,ISSN 0165-1781.
7. Gaur M, Aribandi V, Alambo A, Kursuncu U, Thirunarayan K, Beich J, Pathak J and Sheth A (2021). Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. PLoS One, 16(5), e0250448.
8. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, Sheth A, Welton R and Pathak J (2019). Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention. In: Proceedings of The World Wide Web Conference (WWW'19). New York, USA: Association for Computing Machinery, pp. 514-525.
9. Li, Jun, et al. Suicide risk level prediction and suicide trigger detection: A benchmark dataset. HKIE Transactions Hong Kong Institution of Engineers 29.4 (2022): 268-282.

10. Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O and Bingol HO (2018). Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. The Journal of Medical Internet Research, 20(6), e215.
11. Shing H, Nair S, Zirikly A, Friedenberg M, Daumé H and Resnik P (2018). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. New Orleans, LA: Association for Computational Linguistics, pp. 25-36.
12. Shing H, Resnik P and Oard D (2020). A Prioritization Model for Suicidality Risk Assessment. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 8124-8137.
13. Ji S, Yu CP, Fung S, Pan S, Long G and Cong G (2018). Supervised Learning for Suicidal Ideation Detection in Online User Content. Complexity, 6157249.
14. Linda M. Performance Evaluation of Deep Learning Models on Suicide Ideation Detection of Reddit Posts. The National High School Journal of Science 2024.
15. Tadesse, M.M.; Lin, H.; Xu, B.; Yang, L. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. Algorithms 2020, 13, 7.
16. Michelle Morales, Prajjalita Dey, and Kriti Kohli. 2021. A Comparison of Simple vs. Complex Models for Suicide Risk Assessment. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 99-102.
17. Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, Richard Grucza, and Laura J Bierut. 2016. An analysis of depression, self-harm, and suicidal ideation content on Tumblr. Crisis (2016).
18. Cho SE, Geem ZW, Na KS. Development of a suicide prediction model for the Elderly using Health Screening Data. Int J Environ Res Public Health. 2021;18(19):10150.
19. M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, Discovering shifts to suicidal ideation from mental health content in social media. In Proc. CHI Conf. Hum. Factors Comput. Syst., May 2016, pp. 2098-2110.
20. Vy Nguyen, Chau Pham. Leveraging Large Language Models for Suicide Detection on Social Media with Limited Labels. 2024 IEEE International Conference on Big Data (BigData), pp.8550-8559.
21. Jakub Pokrywka, Jeremi I. Kaczmarek, Edward J. Gorzelañczyk. Evaluating Transformer Models for Suicide Risk Detection on Social Media. 2024 IEEE International Conference on Big Data (BigData), pp.8566-8573.
22. J. Li et al., "Overview of IEEE BigData 2024 Cup Challenges: Suicide Ideation Detection on Social Media," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8532-8540.
23. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.
24. Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, Zettlemoyer Luke and Stoyanov Veselin. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach
25. Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems. Volume 26., 2013.
26. D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo. CLPsych 2016 shared task: Triaging content in online peer-support forums. In Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology, 2016, pp. 118-127.